DOCTEUR EN SCIENCES

Modification de domaines de liaison à la choline en vue de leur utilisation comme étiquette de purification de protéines recombinantes

De Schrevel, Nathalie

Award date:
2005

Awarding institution:
Universite de Namur

FACULTES UNIVERSITAIRES
NOTRE-DAME DE LA PAIX **NAMUR**

FACULTES DES SCIENCES
DEPARTEMENT DE BIOLOGIE

Modification de domaines de liaison à la choline en vue de leur utilisation comme étiquette de purification de protéines recombinantes

Dissertation présentée par
Nathalie De Schrevel
En vue de l'obtention du grade
De Docteur en Sciences

Composition du jury :
Eric Depiereux (Promoteur, FUNDP)
Johan Wouters (FUNDP)
Xavier DeBolle (FUNDP)
Marianne Rooman (ULB, Belgique)
Karin Goraj (GlaxoSmithKline Biologicals, Belgique)

2005
Facultés Universitaires Notre-Dame de la Paix

Rempart de la Vierge, 13
B - 5000 Namur (Belgique)

Imprimé en Belgique
ISBN: 2-87037-503-4
Dépôt légal: D / 2005 / 1881 / 27

Faculté des Sciences
Rue de Bruxelles, 61 B-5000 Namur, Belgique

Modification des domaines de liaison à la choline en vue de leur utilisation comme étiquette de purification de protéines recombinantes

par Nathalie De Schrevel

Résumé

Le but de ce travail consiste à créer un nouveau *tag* de purification d'affinité permettant la purification de protéines recombinantes sur une matrice DEAE-Sépharose Fast Flow.

Dans la nature, certaines protéines de surface de *Streptococcus pneumoniae* sont liées à la paroi bactérienne par des interactions non convalentes faisant intervenir des molécules de choline présentes sur les acides téichoiques et lipotéichoiques. Ces protéines de surface présentent une organisation modulaire avec le domaine catalytique et le domaine de liaison fonctionnant indépendamment l'un de l'autre.

La choline étant un analogue structural du DEAE, l'étude des domaines de liaison à la choline constitue une approche de choix pour concevoir un *tag* de purification présentant une affinité pour le DEAE-Sépharose.

Nous avons plus particulièrement travaillé sur la N-acétyl-L-alanine amidase (LytA) qui dégrade spécifiquement certaines liaisons du peptidoglycan de la paroi de *Streptococcus pneumoniae*. Son domaine de liaison à la choline C-terminal (ClytA) se compose de six motifs répétés imparfaits, constitué chacun d'une vingtaine de résidus.

Deux stratégies ont été développées pour concevoir le *tag* de purification. D'une part, 126 motifs répétés de 19 domaines de liaison à la choline ont été alignés pour définir une séquence consensus. Cette approche a permis de mettre en évidence les résidus importants conservés parmi les motifs répétés. D'autre part, nous avons construit des protéines de fusion portant des fragments du domaine de liaison ClytA de longueur variable. Des expériences de chromatographies sur matrice DEAE-Sépharose nous ont permis d'isoler un petit fragment de ClytA(L234), présentant toujours une affinité spécifique pour le DEAE-Sépharose. Cette affinité est maintenue lorsque le fragment L234 est fusionné à l'extrémité C-terminale d'une autre protéine reporter. Cependant, nos résultats suggèrent que le candidat *tag* L234 est instable et qu'il conduit à l'insolubilisation de la protéine de fusion lors de la production de celle-ci dans *Escherichia coli*. Afin d'améliorer la solubilité/stabilité du fragment L234, nous avons développé trois approches bioinformatiques. Celles-ci ont permis de définir trois groupes de mutations permettant d'améliorer potentiellement la solubilité et/ou la stabilité du fragment L234. Les *tags* mutants ont été construits et fusionnés à l'extrémité C-terminale de la thiorédoxine. Le premier *tag* mutant, EDE-L234, est plus soluble que la version non mutante mais présente une perte d'affinité pour le DEAE-Sépharose. Le second mutant, NG-L234, ne montre pas d'augmentation de solubilité et perd également une partie de son affinité pour la matrice. Le troisième *tag* mutant, V1V2V3-L234, présente une augmentation d'affinité pour le DEAE-Sépharose bien que sa solubilité reste inchangée.

Facultés Universitaires Notre-Dame de la Paix
Faculté des Sciences
Rue de Bruxelles, 61 B-5000 Namur, Belgique

Modification of choline-binding domains for their use as purification tag of recombinant proteins.

by Nathalie De Schrevel

Abstract

The aim of the project was to design a new affinity purification tag for DEAE-Sepharose matrix.

In nature, some surface proteins of *Streptococcus pneumoniae* are attached to the cell wall by non covalent interactions involving choline moieties present on teichoic and lipoteichoic acids. These surface proteins present a modular organization with a catalytic domain and a choline-binding domain which are independant from each other. Since choline is a structural analogue of DEAE, the study of choline-binding domains may be a good starting point to design an affinity purification tag for DEAE-Sepharose. Indeed, when fused to a protein of interest, the complete choline binding domain allows the purification of the fusion protein by affinity chromatography in a single step, using a 2% choline solution for elution.

We have worked on the N-acetylmuramoyl-L-alanine amidase (LytA) which degrades some specific bonds in the peptidoglycan of the cell wall of *Streptococcus pneumoniae*. Its C-terminal choline-binding domain (ClytA) is composed of six imperfect repeats of approximately 20 amino acids each. To design a tag presenting a high affinity for DEAE, we developped two strategies.

First, sequence alignments of 126 repeats from 19 different choline-binding domains allowed the definition of a consensus sequence. This approach highlighted the important residues conserved among all the repeats.

At the same time, we generated different fusion proteins composed of ClytA fragments which varied in size. Purification experiments on DEAE-Sepharose allowed us to isolate a small fragment of ClytA (L234), that still presents affinity for the matrix. The same results were obtained when the L234 fragment was fused to the C-terminal of another reporter protein.

However, our results suggested that L234 tag candidate may be instable and leads to insolubilization of part of the fusion protein in *E. coli* cells.

In an attempt to improve the solubility/stability of the L234 fragment, we developped three bioinformatic approaches. They allowed the definition of three groups of mutations predicted to improve solubility and/or stability of the L234 fragment. These mutant tags were thus constructed and fused to thioredoxin. The first one, called EDE-L234, was more soluble but lost its affinity for DEAE-Sépharose. The second one, NG-L234, was not more soluble and it also lost its affinity. The third mutant tag, called V1V2V3-L234 presented a higher affinity for DEAE than the L234 fragment even if its solubility remained the unchanged.

Remerciements

Ouf…voici la dernière partie de ma thèse à écrire. Je croyais que ce serait la plus facile mais, en me retrouvant devant cette page blanche, je me rends compte que ce n'est pas si évident. Six ans se sont écoulés depuis que j'ai entamé ce travail. En octobre 1999, je croyais naïvement que ma thèse se résumerait à quatre années de travail pour obtenir ce « fichu diplôme » dont j'avais tant besoin pour trouver un job stable après mes expériences africaines. Ces années ont représenté beaucoup plus que cela. Il y a bien évidemment eu ma rencontre avec David, la naissance de notre petite Gaëlle, des moments de joie (enfin un résultat…) et de découragement (je n'y arriverai jamais…) mais aussi la rencontre avec beaucoup de gens chouettes et ouverts. On ne peut par résumer six ans de vie en une ou deux pages, on peut juste remercier les gens de vous avoir permis d'arriver au bout de votre projet.

Je souhaiterais remercier en premier mon promoteur, Eric Depiereux ainsi que Carla Vinals qui ont tous les deux initié ce projet et ont accepté de me faire confiance malgré mon CV « un peu particulier ».

Je remercie également les membres de mon jury, Mesdames Karine Goraj et Marianne Rooman ainsi que Messieurs Xavier De Bolle et Johan Wouters qui ont accepté de lire ce manuscrit et y ont apporté des modifications constructives.

Un merci particulier à Jean-Louis Ruelle, Karine Goraj et Caral Vinals qui, bien qu'ayant des horaires surchargés chez GlaxoSmithKline Biologicals n'ont jamais refusé une réunion pour faire le point sur l'avancement de mes travaux.

Merci à Jean Vandenhaute qui a toujours cru en moi. Sans lui, je n'aurais probablement jamais réintégré l'université ni réalisé de thèse. C'est lui qui, le premier m'a obtenu un poste de neuf mois chez Jean-Jacques Letesson pour « me remettre en selle » après l'Afrique puis qui m'a poussée à accepter ce projet. Même si parfois j'ai douté du bien fondé de cette aventure, je crois que je te dois une fière chandelle.

Ce travail n'aurait pas non plus été possible sans un certain nombre de personnes :

D'abord, il y a Xa, ami de longue date avec lequel j'ai des souvenirs d'études, de voyages, de guindailles, d'intendance à un camp guide et …de thèse. Sans être mon promoteur, tu as passé beaucoup de temps à discuter avec moi, à me suggérer des manips, à critiquer mes résultats. Merci pour ta disponibilité, ton enthousiasme face à la recherche scientifique et ton humeur toujours égale (ça , ça m'épate à chaque fois !!!)
Merci à Cri (Christophe Lambert) qui a eu la (très) lourde tâche d'enseigner à une néophyte la subtilité des alignements de séquences, de l'homology modeling et du threading. Si j'en sais un peu plus aujourd'hui en bioinformatique, c'est certainement grâce à toi.
Merci à Etienne Delaive, le maître des protéines, qui m'a permis de travailler sur la FPLC du laboratoire d'URBC. Je n'ose pas calculer le nombre de fois où il est venu à mon secours « Etienne, l'alarme pression s'est déclenchée», « Etienne, la colonne ne coule plus », « Etienne, le collecteur de fractions ne fonctionne pas ». Je suis incapable de me rappeler le nombre de bugs que j'ai eu sur cette machine par contre, je me rappelle très bien qu'à chaque fois, Etienne était disponible, calme et très efficace.
Merci également à Pierre Cambier qui m'a appris à utiliser le spectrofluorimètre du laboratoire d'URBC.
Un très très grand merci à Fabrice pour la relecture de tous mes résultats et de ma discussion. C'était vraiment très enrichissant pour moi.
La mise en page de ce document ne serait sûrement pas celle que vous voyez aujourd'hui si Sophie Bamps n'avait pas été là. Grâce à elle, j'ai enfin découvert que les logiciels Word et Excell ne servent pas qu'à taper du texte et faire des tableaux. Sophie, je me souviendrai toujours du mot d'ordre principal « l'ordinateur ne fait que ce que tu lui demandes. Donc, tu restes calme et tu réfléchis ». C'est certainement grâce à toi si je n'ai pas jeté mon iMac par la fenêtre la première quinzaine d'août !!
Enfin, il y a Isa. Même si on n'a pas beaucoup discuter « sciences », tu as toujours été là pour me remonter le moral quand je doutais ou pour me tempérer que je « bouillais » (et ça, ce n'est pas toujours de la tarte…). Nos petits briefings matinaux

vont bien me manquer. Maintenant que je ne travaillerai plus aux facs, on a intérêt à se trouver une formule tarif réduit chez un opérateur téléphonique si on ne veut pas se ruiner dans le futur…
Le seul point vraiment négatif de ma thèse est de n'avoir pas pu être intégrée dans une équipe de chercheurs. Cette contrainte découlait de la nature appliquée de mon sujet ainsi que de la clause de confidentialité qui y était liée. Malgré cela, ces six années m'ont permis de découvrir des personnes chouettes et enrichissantes. Aussi, je remercie pour leur présence Anne Titi, Poos, Monique, Rose-Ma, Rose-May, Zette, Bassam, Calou, Manu Gillot, Val, Ismaëlle, Christine, Valérie, grande Marie, petite Marie, Chantal, Godi, Cri, Benjamin, Régis, Benoît, Amélie, Garçon, Flore, JYP, Allan, Lionel, Etienne, Godi, Aïko, Marie-Ange, Max, ainsi que la bande des « petits jeunes » que je connais moins bien (Sophie, Johann, Julien, Nico, Virginie, Alex, Richard, Jonathan,…). Je vous souhaite bon vent à tous.
Puis, il y a tous ceux qui ne naviguent pas dans les sphères de la recherche scientifique mais qui m'ont apporté un précieux soutien. Mes copines de chez MSF : Isa, Maureen, Katelijn, Dominique, Edith,…et bien sûr ma famille.
Un merci tout particulier à mes parents pour leur soutien indéfectible dans toutes les étapes de ma vie. Merci à Marie, Christophe, Jean-Philippe et Nubia pour leurs conversations réconfortantes et pour les bons petits soupers familiaux.
Enfin un énorme merci à mes trois amours, David, Laura et Gaëlle pour leur patience et leur présence.

LISTE DES ABREVIATIONS

AA	acide aminé
ADN	acide désoxyribonucléique
Å	Angström
$Å^3$	Angström cube
°C	degré centigrade
CBP	"*Choline-Binding Protein*"
CAT	Chloramphénicol-Acétyl-Transférase.
Da	Dalton
DEAE	Diéthylaminoéthyle
h	heure
GBP	"*Glucan Binding Protein*"
g	gramme
GTF	Glucosyltransférase
IPTG	isopropyl-β-D-thiogalactoside
l	litre
LTA	"Lipoteichoic Acid" ou acide lipotéichoïque
LPS	lipopolysaccharide
M	molaire
mDpm	acide m-diamino-pimélique
mg	milligramme
ml	millilitre
mM	millimolaire
µg	microgramme
µl	microlitre
NAG	N-acétylglucosamine
NAM	acide N-acétylmuramique
nm	nanomètre
PBS	"*Phosphate Buffer Saline*"
PDB	"*Protein Data Bank*"
pI	point isoélectrique
rpm	révolution par minute
TA	"*Teichoic Acid*" ou acide téichoïque
u.a.	unité arbitraire

TABLE DES MATIERES

1. L'étiquetage et la purification des protéines 1
1.1. L'étiquetage des protéines : un outil intéressant 1
1.2. La purification des protéines par chromatographie 2
1.2.1. La chromatographie par échange d'ion 3
1.2.2. La chromatographie d'affinité 4
1.3. Le DEAE-Sépharose et son analogue structural, la choline 7
2. La paroi et les protéines de surface de Streptococcus pneumoniae 8
2.1. La structure générale de la paroi des bactéries Gram+ 8
2.1.1. Structure générale du peptidoglycan 9
2.1.2. Les acides téichoïques et lipotéichoïques des bactéries Gram+ 11
2.1.3. Les acides téichuroniques et les lipoglycans. 15
2.2. Les protéines de surface des bactéries Gram+ 16
2.2.1. Présentation générale des protéines de surface des bactéries Gram+ 16
2.2.2. Les protéines liées de façon covalente au peptidoglycan par un motif LPXTG 19
2.2.3. Les protéines liées aux lipides de la membrane cytoplasmique 20
2.2.4. Les protéines de surface possédant des modules GW 20
2.2.5. Les protéines de surface ancrées dans la membrane plasmique 20
2.2.6. Les protéines liées de façon non covalente par des interactions électrostatiques 21
2.3. La structure chimique de la paroi de *Streptococcus pneumoniae* 22
2.4. La choline pariétale chez *Streptococcus pneumoniae* 24
2.5. Les protéines de surface possédant un domaine de liaison à la choline 28
2.5.1. Les 'choline-binding proteins'(CBP) 29
2.5.2. La structure des domaines de liaison à la choline 31
2.5.3. L'amidase LytA de Streptococcus pneumoniae 33
3. Stratégies de conception d'un tag de purification 39
3.1. Trois approches envisageables pour créer un tag 39
3.2. Les outils bioinformatique utilisables pour la conception d'un tag de purification 39
3.3. L'approche bioinformatique dans le cadre de la création d'un tag de purification par affinité pour le DEAE-Sépharose 42
PARTIE I : ANALYSES BIOINFORMATIQUES DES DOMAINES DE LIAISON À LA CHOLINE 45
I.1 Analyse des caractéristiques physico-chimiques des séquences des domaines de liaison à la choline 45
I.2. Etablissement d'un consensus des séquences répétées composant les domaines de liaison à la choline 46
I.2.1. Définition des motifs répétés 46
I.2.2. Alignements multiples sur les 133 motifs répétés 47
I.2.3. Alignements multiples sur les 126 repeats 47
I.2.4. Classification des motifs répétés 51
I.2.4.1. Classification des repeats par degré de similarité 51
I.2.4.2. Classification des motifs en fonction de leur position dans le domaine de liaison 52
I.2.4.3. Comparaison des consensus établis et établissement d'un consensus final 53
I.3. Prédiction des structures secondaires et recherche de structures tri-dimentionnelles connues, proches des domaines de liaison à la choline 55
I.3.1. Prédiction de structures secondaires sur l'entièreté des domaines de liaison à la choline et sur le consensus final 55
I.3.2. Recherche de structures tridimentionnelles connues, proches des domaines de liaison à la choline 56
I.4. Les structures des domaines de liaison à la choline ClytA et CPLI 61

I.4.1. Présentation générale de la structure du domaine de liaison partiel de l'amidase LytA de *Streptococcus pneumoniae* 61
I.4.2. Présentation générale de la structure du lysosyme CPL1 du phage Cp-1 67
I.4.3. Analyse du consensus de séquences établi par confrontation avec les structures cristallographiques de ClytA et CPL1 69
I.5. Conclusions de l'analyse bioinformatique *73*
PARTIE II : SELECTION D'UN FRAGMENT DU DOMAINE DE LIAISON CLYTA 77
II.1. Dissection du domaine de liaison ClytA *77*
II.1.1. Construction des protéines de fusion thiorédoxine-domaine de liaison ClytA tronqué 77
II.1.2. Estimation indirecte de l'état de repliement des protéines de fusion 79
II.2. Analyse de l'affinité des fragments du domaine de liaison pour le DEAE-Sépharose Fast Flow *82*
II.2.1. Estimation de l'affinité pour le DEAE-Sépharose des protéines de fusion pré-purifiées 82
II.2.2. Estimation de l'affinité pour le DEAE-Sépharose Fast Flow de la thiorédoxine-L234 présente dans la fraction soluble d'*E. coli* 85
II.3. Validation du fragment L234 en tant que tag de purification *87*
II.3.1. Construction de la protéine de fusion MiaA-L234 87
II.3.2. Estimation de l'affinité de la protéine de fusion MiaA-L234 pour le DEAE-Sépharose Fast Flow 88
PARTIE III : OPTIMALISATION DES CONDITIONS DE PRODUCTION DE LA PROTEINE THIOREDOXINE-L234 91
PARTIE IV : DEFINITION DE MUTATIONS PONCTUELLES EN VUE D'AMELIORER LA SOLUBILITE / STABILITE DU *TAG* ET ANALYSE DES MUTANTS 98
IV.1. Définition de mutations ponctuelles *98*
IV.1.1. Recherche de résidus hydrophobes exposés au solvant suite au raccourcissement du domaine de liaison ClytA 98
IV.1.2. Sélection de mutations par comparaison de la séquence des repeats 2, 3 et 4 au consensus de séquence d'un repeat 100
IV.1.3. Sélections de mutations stabilisantes par l'algorithme PoPMuSiC 103
IV.1.4. Définition de trois mutants 105
IV.2. Construction et analyse des mutants thiorédoxine-EDE-L234, thiorédoxine-NG-L234 et thiorédoxine-V1V2V3-L234 ..... *107*
IV.2.1. Construction des protéines de fusion mutantes 107
IV.2.2. Estimation de l'affinité pour le DEAE-sépharose Fast Flow des candidats tags mutants pré-purifiés 109
IV.2.3. Estimation de l'affinité pour le DEAE-Sépharose des protéines thiorédoxine-EDE-L234 et thiorédoxine-V1V2V3-L234 présentes dans la fraction soluble d'*E. coli*. 111
IV. 3. Estimation de la solubilité des protéines de fusion thiorédoxine-EDE-L234 et thiorédoxine-V1V2V3-L234 ..... *114*
1. Souches bactériennes et vecteurs plasmidiques *133*
1.1. Les souches d'*Escherichia coli* 133
1.1.1. Escherichia coli DH10B 133
1.1.2. Escherichi coli BL21(λDE3) (Novagen) 134
1.2. Vecteurs plasmidiques 134
1.2.1. Vecteur pBluescript SK(+) (Stratagene) 134
1.2.2. Vecteur pET15b (Novagen) 135
2. Amplification et clonage de la thiorédoxine et des fragments de domaine ClytA dans le vecteur pET15b *135*

4. Surexpression de la thiorédoxine et des protéines de fusion thiorédoxine-domaine tronqué *138*
4.1. Principe de la surexpression de protéines recombinantes avec le système pET 138
4.2. Conditions standards de surexpression de protéines recombinantes avec le système pET et préparation des fractions soluble et insoluble 138
4.3. Protocole de surexpression pour l'optimisation de la température de production 139
4.4. Protocole de surexpression pour l'optimisation de la concentration en IPTG .. 140
4.5. Protocole de surexpression pour l'optimisation de la composition du milieu de culture 140
4.6. Protocole de surexpression de la thiorédoxine-L234 et des protéines mutantes thiorédoxine-EDE-L234, thiorédoxine-NG-L234 et thiorédoxine-V1V2V3-L234 141
5. Purification sur colonne de chélation : chromatographie d'affinité par immobilisation de métaux (IMAC) *142*
6. Dosage protéique *142*
7. Précipitation des protéines au sulfate d'ammonium *142*
8. Estimation de l'affinité des protéines d'intérêt pour le DEAE-Sépharose ... *143*
8.1. Estimation de l'affinité des protéines de fusion thiorédoxine-fragments de domaine lorsqu'elles sont pré-purifiées ou présentes dans la fraction soluble d'*E. coli* 143
8.2 Mise en évidence des différences d'affinité pour le DEAE-Sépharose des protéines purifiées thiorédoxine-L234, thiorédoxine-EDE-L234, thiorédoxine-NG-L234 et thiorédoxine-V1V2V3-L234 144
9. Dosage des bandes protéiques en SDS-PAGE par "scanning" densitométrie après coloration du gel au Bleu de Coomasie *144*
10. Analyses en Western blot des fractions protéiques *144*
11. Mesure de la fluorescence des protéines de fusion thiorédoxine-fragment du domaine de liaison ClytA *145*
12. Programmes bioinformatiques *146*
12.1. L'algorithme PSI-Blast 146
12.2. L'algorithme SAPS 146
12.3. Programme d'alignement multiple ClustalW 146
12.4. Match-Box 147
12.5. Dialign 2 148
12.6. Blockmaker 148
12.7. PSI-PRED 149
12.8. Les serveurs de pseudo-*threading* 150
12.9. Les serveurs de *threading* 150

ANNEXE 1 ..... **151**

ANNEXE 2 ..... **159**

ANNEXE 3 ..... **165**

ANNEXE 4 ..... **167**

BIBLIOGRAPHIE **169**

INTRODUCTION

1. L'étiquetage et la purification des protéines

1.1. L'étiquetage des protéines : un outil intéressant

La biologie, considérée il y a encore peu de temps comme l'étude de la faune et de la flore, a connu un essor considérable ces trente dernières années. Avec le développement de techniques telles que celle de l'ADN recombinant, les champs d'étude de la biologie se sont développés. On parle actuellement de séquençage de génomes, de protéomes, d'interactomes… Si ces nouvelles stratégies permettent une approche plus globale dans la compréhension du vivant, elles nécessitent également le développement de techniques de plus en plus sophistiquées et d'outils de plus en plus performants. Un de ces outils consiste à marquer (ou étiqueter) des protéines d'intérêt.

En recherche fondamentale, les applications de l'étiquetage des protéines sont très vastes. Elles englobent par exemple, la localisation et les mouvements cellulaires, la mise en évidence d'interactions avec d'autres protéines et la topologie des protéines de membrane (Jarvik and Telmer, 1998).

Une autre application importante de l'étiquetage des protéines concerne leur purification par chromatographie. En effet, sans que nous le soupçonnions, nombre de protéines purifiées interviennent dans la vie courante (Kirk et al., 2002). Le secteur agro-alimentaire en est un gros utilisateur puisque de nombreuses enzymes sont utilisées dans la boulangerie, la fabrication des jus de fruits, la brasserie,… D'autres secteurs tels que l'industrie du textile, la fabrication du papier, l'industrie des détergents et produits lessiviels et l'industrie pharmaceutique font également un usage intensif des protéines purifiées. Outre l'aspect économique lié à l'augmentation de la vitesse réactionnelle des réactions biochimiques, l'utilisation de protéines purifiées représente un enjeu écologique important. En effet, dans de nombreux procédés de fabrication, elles remplacent avantageusement les produits chimiques de synthèse, permettant ainsi de faire de réels progrès dans la réduction des déchets émanant de ces procédés, grâce à la biodégradabilité et à une moindre consommation d'énergie. Etant donné qu'elles ont une action plus spécifique que les produits chimiques de synthèse, les procédés qui les utilisent ont moins de réactions secondaires et de sous-produits résiduaires et donnent des produits de meilleure qualité tout en diminuant la probabilité de pollution Enfin, d'un point de vue médical, l'utilisation de protéines recombinantes purifiées évite les réactions indésirables dues aux contaminants cellulaires et microbiens tels que le lipide A d'*Escherichia coli*.

Cette brève introduction met en évidence l'importance sous-jacente des procédés de purification des protéines et l'intérêt toujours croissant du monde industriel pour la mise au point de nouveaux procédés de

purification. C'est dans cette thématique que s'inscrit le sujet de notre travail.

1.2. La purification des protéines par chromatographie

La purification d'une protéine d'intérêt peut faire intervenir des techniques très diverses comme la précipitation suivie d'une centrifugation, l'extraction par certains solvants, l'électrophorèse ou encore la chromatographie. Bien que ces différentes techniques soient complémentaires, nous nous intéresserons principalement, dans cette introduction, à décrire des techniques de purification par chromatographie.
La séparation par chromatographie dépend de la partition préférentielle des protéines entre une phase stationnaire (l'adsorbant) et une phase mobile.
Les différents adsorbants utilisés exploitent des propriétés variées de la protéine d'intérêt. Parmi les techniques les plus couramment utilisées, nous pouvons citer :

- la perméation sur gel qui se base sur la taille et la forme des protéines,
- la chromatographie par échange d'ions faisant intervenir la charge nette et la répartition des groupements chargés à la surface des protéines,
- la "*chromatofocusing*" dans laquelle les protéines se séparent en fonction de leur point isoélectrique,
- la chromatographie par interactions hydrophobes et la chromatographie en phase inverse faisant intervenir l'hydrophobicité des protéines comme critère de séparation,
- la chromatographie d'affinité exploitant les affinités biospécifiques de certaines protéines pour des ligands.

Pour obtenir un bon degré de purification de la protéine d'intérêt, au départ d'un extrait biologique brut, il faut souvent combiner plusieurs techniques de chromatographie.
Bien que chaque technique chromatographique ait ses avantages et inconvénients, nous ne décrirons, dans la suite de ce chapitre, que les principes de la chromatographie par échange d'ions et chromatographie d'affinité, indispensables pour la compréhension de notre travail.

1.2.1. La chromatographie par échange d'ion

De façon générale, le principe de cette chromatographie se base sur l'adsorption réversible de molécules solubles chargées, à des groupes ioniques de charge opposée, immobilisés sur un support inerte (Karlsson et al., 1989) La purification est possible grâce au fait que des substances diverses présentent des degrés d'interactions variables avec les groupements échangeurs ioniques suite à une différence dans leur charge nette et dans la densité et répartition de ces charges à leur surface.
Donc, lorsque l'échantillon à purifier est appliqué sur la colonne, les molécules solubles, portant la charge appropriée, se lient de façon réversible à la matrice. Puis, elles sont éluées de la colonne en appliquant des conditions défavorables pour leurs liaisons ioniques avec le support. Ces conditions sont soit une augmentation de la force ionique du tampon d'élution soit un changement de son pH. Dans le cas d'un gradient de sels, l'élution des molécules solubles est fonction de la force de leurs liaisons ioniques avec le support, les molécules les plus faiblement liées s'éluant en premier (figure 1).

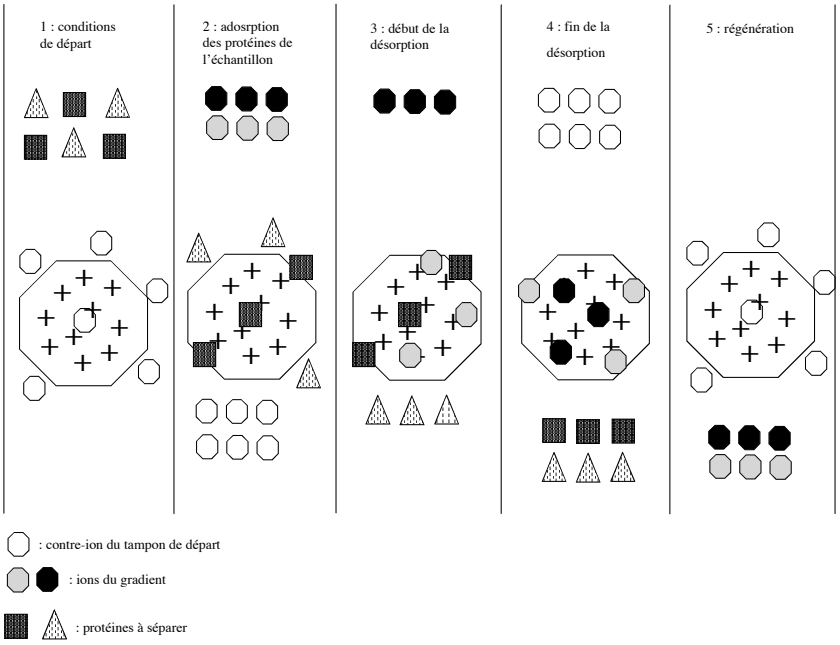


Figure 1 Illustration du principe de séparation des protéines par échange d'ions (modifié de "Ion Exchange Chromatography, Principle and Methods", Pharmacia Biotech., ISBN 91 970490-3-4).
L'échange d'ions est probablement une des techniques chromatographiques les plus utilisées pour séparer et purifier des protéines, polypeptides, acides

nucléiques, et autres molécules chargées. Parmi les raisons de son succès, on peut citer sa large applicabilité, son haut taux de résolution, sa grande capacité de charge (voir matériel et méthodes, paragraphe 8), la simplicité et le contrôle aisé de la méthode. De plus, elle permet de purifier des molécules dans des conditions non-dénaturantes, proches des conditions physiologiques.
Enfin, d'un point de vue industriel, l'utilisation de ce type de gel chromatographique est bien caractérisé et relativement peu onéreux.

1.2.2. La chromatographie d'affinité

La chromatographie d'affinité fait intervenir une interaction biospécifique entre la protéine d'intérêt et un ligand (inhibiteurs, récepteurs, anticorps, sucres, …) et permet, assez fréquemment, une bonne purification en une seule étape.
Lorsque l'on veut purifier une protéine par affinité, deux stratégies sont possibles :
- créer une colonne d'affinité spécifique de la protéine d'intérêt en greffant, par exemple, un anticorps ou un inhibiteur spécifique de cette protéine sur un support inerte (Pedersen et al., 2004) (Nord et al., 2000).
- marquer la protéine à l'aide d'un *tag* universel permettant la purification en une étape sur un type de support bien défini (Nilsson et al., 1997) (Terpe, 2003).

Les deux approches possèdent leurs avantages et inconvénients. Le plus grand avantage du *tag* de purification est sa facilité d'utilisation. La méthode de purification étant déjà mise au point, son utilisation est rapide et souvent peu coûteuse. De plus, elle ne nécessite pas l'isolement préalable ni la caractérisation des propriétés biochimiques de la protéine d'intérêt. Dans de nombreux cas, le *tag* aide à la solubilité, la stabilité et le repliement de la protéine d'intérêt dans un hôte particulier. Par contre, il est parfois nécessaire de se débarrasser du *tag* après purification. Cette nécessité dépend de l'utilisation ultérieure de la protéine d'intérêt (utilisation dans l'industrie pharmaceutique, cristallographie, études NMR…). Dans un tel cas de figure, l'enlèvement du *tag* se fait souvent par digestion enzymatique à un site protéase-spécifique situé entre le *tag* et la protéine d'intérêt. L'efficience du clivage protéolytique n'est pas toujours optimale et il est souvent nécessaire de re-séparer par chromatographie le *tag* clivé de la protéine d'intérêt, diminuant de ce fait le rendement de purification. De plus, la fusion d'une protéine d'intérêt avec un *tag* ne garantit pas toujours que la protéine adopte sa conformation native suite à l'influence du *tag* sur la structure tri-dimensionnelle de la protéine de fusion. Il peut s'en suivre une perte de l'activité spécifique de la protéine d'intérêt. Bien que réels, ces désavantages ne sont pas très courants, comme le prouvent de nombreuses

données de la littérature relatant le succès de purification de protéines d'intérêt par des *tag*s universels (Einhauer and Jungbauer, 2001) (Smith and Johnson, 1988) (Uhlen et al., 1983).
En conclusion, l'utilisation de *tags* reste une voie royale de purification des protéines car elle est rapide et demande peu de mise au point.
Les *tags* les plus fréquemment utilisés trouvent souvent leur origine dans la nature. En effet, le monde vivant présente un choix extraordinaire d'exemples d'interactions spécifiques entre une protéine et un ligand. C'est donc, notamment, par l'observation de la nature que le biologiste peut conceptualiser de nouveaux systèmes de purification par chromatographie d'affinité. Une autre approche consiste à baser la conception du *tag* uniquement sur des propriétés physico-chimiques telles que les charges électriques, l'hydrophobicité … Cette approche, plus chère aux chimistes et physiciens, a donné naissance à un certain nombre de systèmes de purification tels que le *tag* poly-arginines (Smith et al., 1984) ou poly-phénylalanines (Persson et al., 1988).
Dans la littérature, on classe souvent les *tags* de purification les plus fréquemment utilisés en fonction de leur degré d'organisation : protéine complète, domaine protéique, peptides dérivés d'un domaine protéique, combinaison de différents *tags*, etc…
Les *tags* commerciaux les plus fréquemment utilisés sont repris au tableau 1.
Un *tag* dédicacé à la purification par affinité doit, si possible, posséder un certain nombre de qualités dont la principale est sa capacité à adopter sa structure tertiaire sans interférer avec la structure de la protéine d'intérêt (Einhauer and Jungbauer, 2001). Cette propriété, bien que non prédictible, est la première condition pour sélectionner un candidat. Sa bonne solubilité et sa stabilité protéolytique sont également des facteurs importants (Graslund et al., 2002). D'un point de vue plus technique, le *tag* doit pouvoir lier sa cible de façon réversible dans des conditions expérimentales douces et peu coûteuses tout en gardant une bonne sélectivité et affinité. Le ligand du *tag* de purification doit être greffable sur un support et supporter plusieurs cycles de nettoyage de la colonne (Nord et al., 2000). Si la protéine d'intérêt est de taille importante, il est parfois préférable d'utiliser un *tag* de petite taille.afin de ne pas influencer le rendement de production. Enfin, d'autres qualités telles que la compatibilité avec la sécrétion et l'absence de cystéines pouvant interférer avec la formation de ponts disulfures de la protéines d'intérêt sont des facteurs non négligeables (di Guan et al., 1988).

A : Matrices and elution conditions of affinity tags

Affinity tag	matrix	Elution condition
Poly-Arg	Cation-exchange resin	NaCl linear gradient from 0 to 400 mM at alkaline pH>8
Poly-His	Ni2+-NTA, Co2+-CMA (Talon)	imidazole 20-250 mM or low pH
FLAG	Anti-FLAG monoclonal antibody	pH 3.0 or 2-5 mM EDTA
Strep-tag II	Strep-Tactin (modified streptavidin)	2.5 mM desthiobiotin
c-myc	Monoclonal antibody	low pH
S	S-fragment of RNaseA	3M guanidine thiocyanate, 0.2 M citrate pH2, 3M magnesium chloride
HAT (natural histidine affinity tag)	Co2+-CMA (Talon)	150 mM imidazole or low pH
Calmodulin-binding peptide	Calmodulin	EGTA or EGTA with 1M NaCl
Cellulose-binding domain	Cellulose	Family I : guanidine HCl or urea>4M Family II/III : ethylene glycol
SBP	streptavidin	2 mM Biotin
Chitin-binding domain	Chitin	Fused with intein : 30-50 mM dithiothreitol, b-mercaptoethanol or cysteine
Glutathione S-transferase	Glutathione	5-10 mM reduced glutathione
Maltose-binding protein	Cross-linked amylose	10 mM maltose

B : Sequence and size of affinity tags

Tag	Residues	Sequence	Size (kDa)
Poly-Arg	5-6 (usually 5)	RRRRR	0,8
Poly-His	2-10 (usually 6)	HHHHHH	0,84
FLAG	8	DYKDDDDK	1,01
Strep-tag II	8	WSHPQFEK	1,06
c-myc	11	EQKLISEEDL	1,2
S	15	KETAAAKFERQHMDS	1,75
HAT-	19	KDHLIHNVHKEFHAHAHNK	2,31
3X FLAG	22	DYKDHDGDYKDHDIDYKDDDDK	2,73
Calmodulin-binding peptide	26	KRRWKKNFIAVSAANRFKKISSSGAL	2,96
Cellulose-binding domain	27-189	Domains	3,00-20,00
SBP	38	MDEKTTGWRGGHVVEGLAGELEQLRARLEHHPQGQREP	4,03
Chitin-binding domain	51	TNPGVSAWQVNTAYTAGQLVTYNGKTYKCLQPHTSLAGWEPSNVPALWQLQ	5,59
Glutathione S-transferase	211	Protein	26
Maltose-binding protein	396	Protein	40

Tableau 1 Liste des *tag* s fréquemment utilisés dans l'industrie et la recherche fondamentale (Terpe, 2003)

1.3. Le DEAE-Sépharose et son analogue structural, la choline

Comme nous l'avons vu au point 1.2.1, si la chromatographie par échange d'ions présente de nombreux avantages en termes d'applicabilité, capacité de charge, résolution et facilité, elle nécessite cependant une mise au point pour chaque nouvelle protéine à purifier. Par contre, en utilisant un *tag* de purification en chromatographie d'affinité, la purification d'une protéine peut se faire en une étape, sans beaucoup de mise au point préalable. Allier les avantages de ces deux techniques chromatographiques présente donc un intérêt particulier et constitue l'objet de ce travail.
Le support chromatographique que nous avons choisi, en partenariat avec la firme GalxoSmithKline Biologicals, est le diéthylaminoéthyle-Sépharose (DEAE-Sépharose). Caractérisé par une amine tertiaire, ce support est très fréquemment utilisé dans la chromatographie par échange d'ions et est considéré comme un échangeur d'anion faible, l'état de protonation de l'amine tertiaire étant fonction du pH appliqué.
Comme nous l'avons signalé au paragraphe précédent, les *tags* les plus fréquemment utilisés trouvent souvent leur origine dans la nature.
Dans le cadre de notre travail, la conception d'un nouveau *tag* va trouver son origine dans l'existence de domaines de liaison à la choline, présents chez des protéines de surface de certaines bactéries à Gram+, dont *Streptococcus pneumoniae*, et chez les protéines de certains phages. La principale caractéristique de la choline, du point de vue de la conception d'un *tag*, est son analogie structurale avec le DEAE (figure 2). Des expériences de purification de protéines de fusion portant un domaine de liaison à la choline montrent qu'il est possible de purifier ces protéines sur un support chromatographique de type DEAE-cellulose, en utilisant la choline comme molécule d'élution (Sanchez-Puelles et al., 1992) (Ortega et al., 1992).

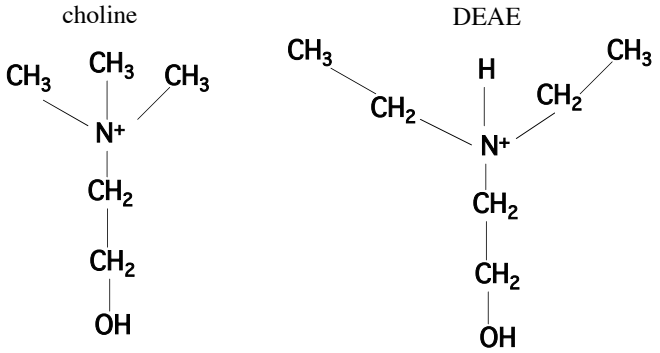


Figure 2 Représentation de la choline et du DEAE.
Dans une matrice DEAE-Sépharose, les groupements DEAE sont liés à la matrice de sépharose par des liaisons éthers entre l'hydroxyle du DEAE et les unités monosaccharidiques.

Avant de passer à une description plus détaillée de ces domaines, nous donnerons d'abord un bref rappel de la composition de la paroi des bactéries Gram+ ainsi que de protéines de surface de ces bactéries puis, nous présenterons la structure particulière de la paroi de *Streptococcus pneumoniae*. Enfin, nous décrirons les domaines de liaison à la choline, et, plus particulièrement celui de l'amidase LytA de *Streptococcus pneumoniae* qui nous servira de point de départ pour la conception d'un *tag*.

2. La paroi et les protéines de surface de *Streptococcus pneumoniae*

2.1. La structure générale de la paroi des bactéries Gram+

Si l'enveloppe des bactéries Gram – est constituée uniquement de deux membranes séparées par une fine couche de muréine, la paroi des bactéries Gram+ est composée d'une bicouche de phospholipides (membrane cytoplasmique), d'un espace périplasmique très étroit et d'une paroi (figure 3).

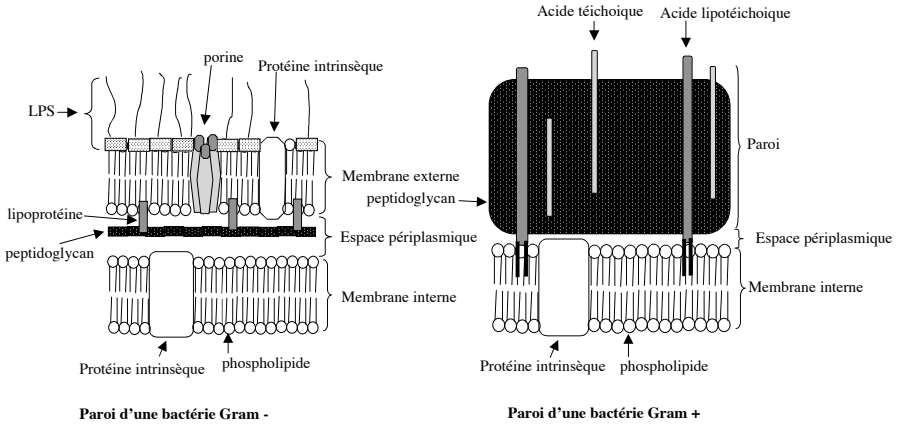


Figure 3 Schémas de la paroi d'une bactérie à Gram négatif et de celle d'une bactérie Gram positif
(modifié de
http://pedagogie.acmontpellier.fr/Disciplines/sti/biotechn/microbio.html).

A l'extérieur de cette enveloppe, on trouve, chez beaucoup de bactéries, un manteau fibreux externe, appelé le glycocalyx. Si ce dernier présente plutôt une structure gélatineuse associée étroitement à l'enveloppe, on parle de capsule. Par contre, si la glycocalyx semble peu solidaire de l'enveloppe et désorganisé, on parle plutôt de couche visqueuse. Composé de polysaccharides et/ou de polypeptides, on attribue au glycocalyx des fonctions dans l'adhérence à des structures de l'environnement, dans la protection contre la phagocytose, dans la résistance au dessèchement, dans l'accumulation de certains nutriments, dans l'accès de macromolécules et

d'ions et dans le stockage de déchets du métabolisme (Ingraham and Ingraham, 1995) (Neuhaus and Baddiley, 2003).

Chez les bactéries Gram+, la paroi est le support de toute une série de molécules et remplit de multiples fonctions dont beaucoup sont critiques pour la viabilité de la cellule. La fonction première de la paroi consiste à fournir un exosquelette rigide protégeant contre la lyse osmotique et mécanique (Navarre and Schneewind, 1999).
La paroi cellulaire bactérienne des bactéries Gram+ est constituée d'une macromolécule de peptidoglycan (encore appelé muréine), sur laquelle viennent se greffer des polymères secondaires anioniques tels que les acides téichoïques et lipotéichoïques.

2.1.1. Structure générale du peptidoglycan

Le peptidoglycan consiste en la répétition d'un disaccharide universellement conservé chez les bactéries : N-acétyl-glucosamine (NAG) - acide N-acétyl-muramique (NAM). Les disaccharides sont reliés entre-eux par des liaisons glucosidiques β 1-4 (figure 4). Le nombre d'unités disaccharidiques constituant un brin de glucan est très variable d'une espèce bactérienne à l'autre (5 à 30 sous-unités). Lors de sa synthèse, l'acide N-acétyl-muramique porte un pentapeptide latéral de base composé séquentiellement de L-alanine, d'acide D-glutamique, d'un acide diaminé (l'acide méso-diamino-pimélique (m-Dpm) ou la lysine) et de deux résidus D-alanine. Dans la paroi, les brins de glucan sont reliés entre-eux par des ponts peptidiques dont la nature est le principal facteur de diversité parmi les peptidoglycans actuellement répertoriés (Schleifer and Kandler, 1972). Lors de la formation de ces ponts peptidiques, la dernière D-alanine est clivée, transformant le pentapeptide en tétrapeptide. Chez certaines bactéries Gram+, la formation des ponts peptidiques entre les brins de glucan se produit par liaison directe du groupement amine de l'acide m-diamino-pimélique (mDpm ,intermédiaire dans la voie de biosynthèse de la lysine) ou de la lysine situés en position 3 du pentapeptide avec le groupement carbonyle de l'avant-dernière D-alanine d'un autre pentapeptide.

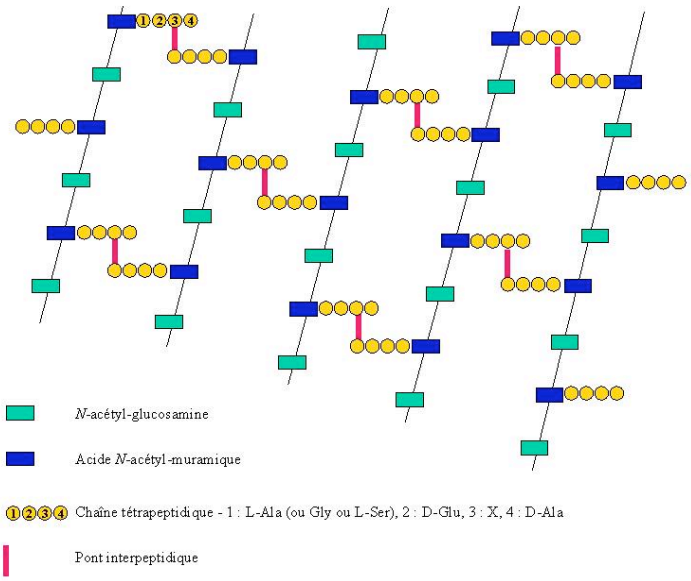


Figure 4 Structure schématique du peptidoglycan (http://www.bacterio.cict.fr/bacdico/bacteriogene/schpeptidoglycane.html).

Par contre, chez d'autres bactéries Gram+, la liaison des peptides latéraux de deux brins de glucan se fait par insertion d'un pont peptidique supplémentaire entre l'acide aminé en position 3 d'un peptide et l'avant-dernière D-alanine d'un autre peptide. Ce pont peut être constitué d'acides aminés tels que la glycine, la thréonine, la sérine ou encore l'acide aspartique. (Madigan et al., 2003). C'est notamment le cas chez *Staphylococcus aureus* où la liaison peptidique entre les tétrapeptides de deux brins de glucan fait intervenir un peptide composé de 5 glycines (figure 5).

Le pourcentage de liens peptidiques entre les brins de glucan présente aussi une grande variabilité d'une espèce bactérienne à l'autre. Il est de l'ordre de 25% chez *Escherichia coli* et est souvent plus important chez les bactéries Gram+.

Le rôle de protection contre la lyse mécanique attribué à la paroi est principalement rempli par le peptidoglycan.

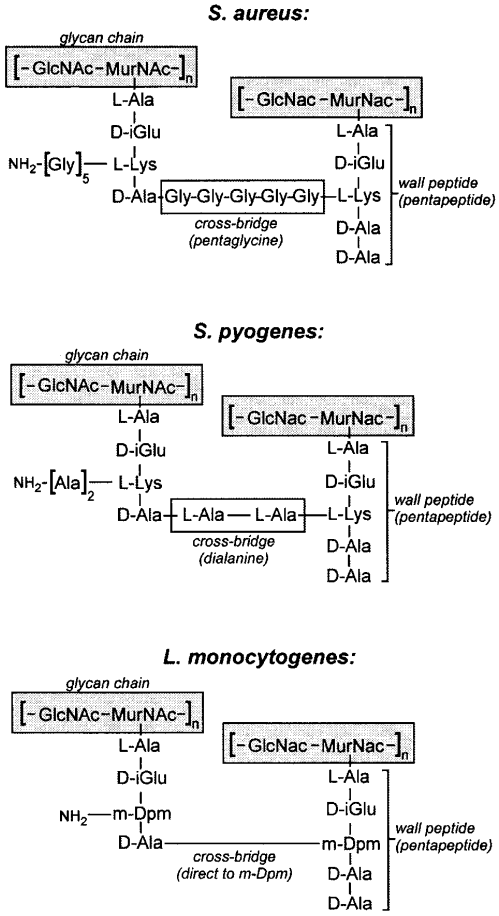


Figure 5 Exemples de ponts peptidiques reliant deux brins de glucan (Navarre and Schneewind, 1999).

2.1.2. Les acides téichoïques et lipotéichoïques des bactéries Gram+

Les acides téichoïques et lipotéichoïques sont définis comme des polymères présentant des groupements phosphodiesters, des polyols et/ou des résidus sucrés, et souvent, mais pas toujours, des résidus estérifiés de D-alanine (figure 6) (Ward, 1981). Ces polymères sont chargés négativement.
Chez les bactéries Gram+, les acides téichoïque (TA) et lipotéichoïque (LTA) ont généralement une structure chimique différente, bien qu'il existe des exceptions (paragraphe 2.3). Ces deux polymères sont distribués sur l'entièreté du peptidoglycan et leur charge globale est négative (Navarre and Schneewind, 1999).

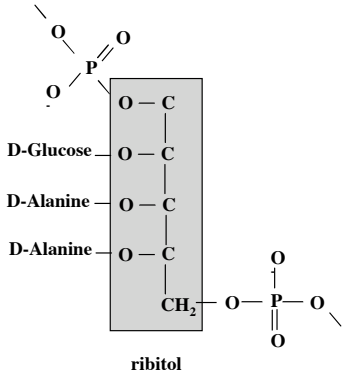


Figure 6 Représentation schématique de la structure d'un acide téichoïque polyribitol phosphate, substitué par des D-Alanine et des D-glucose. La structure du ribitol est surlignée en gris.

Les acides téichoïques sont attachés au peptidoglycan par formation d'un lien covalent entre le dernier groupement phosphate et le C_6 des acides N-acétyl-muramique. Ils consistent généralement en un polymère de polyribitol phosphate ou polyglycérol phosphate, souvent glycosylés ou estérifiés par des acides aminés à divers degrés. La figure 7 donne un aperçu de la diversité de structure des acides téichoïques.

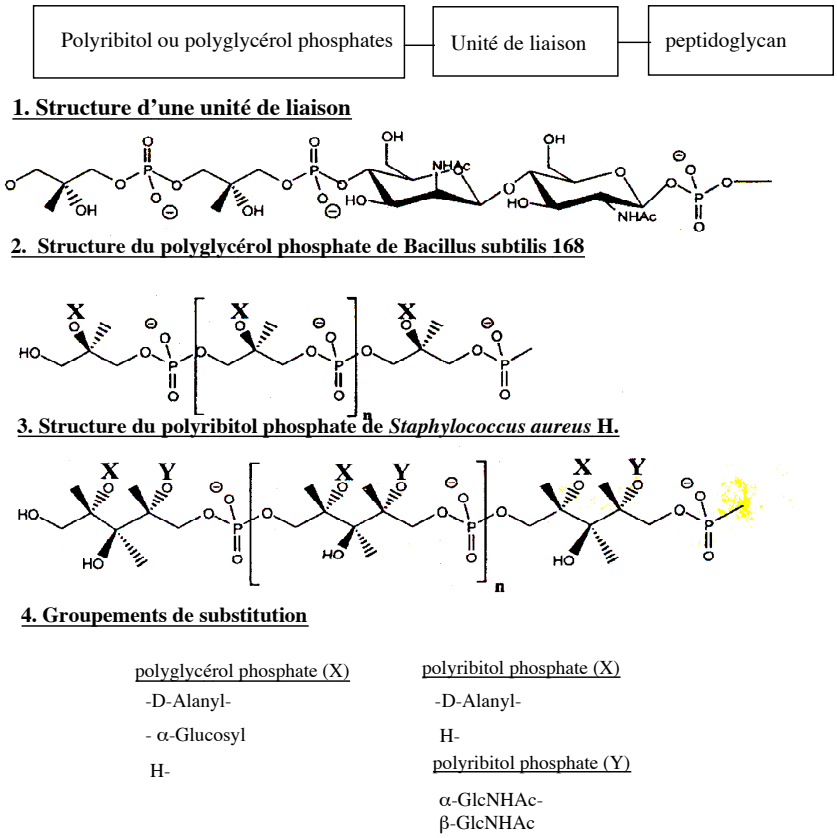


Figure 7 Structures des acides téichoïques de *Bacillus subtilis* 168 et de *Staphylococcus aureus* H (modifié de (Neuhaus and Baddiley, 2003)).

Contrairement à l'acide téichoïque, l'acide lipotéichoïque n'est pas attaché de façon covalente au peptidoglycan mais est inséré dans le feuillet externe de la membrane cytoplasmique par une ancre lipidique. Ce polymère s'étend à travers toute la paroi jusqu'à la surface de la bactérie. La faible hétérogénéité observée parmi les acides lipotéichoïques de diverses bactéries provient de la composition en acide gras, du type de glycosilation et du degré de substitution par ces résidus glycosylés, de la longueur de la chaîne et du degré de D-alanilation (figure 8).

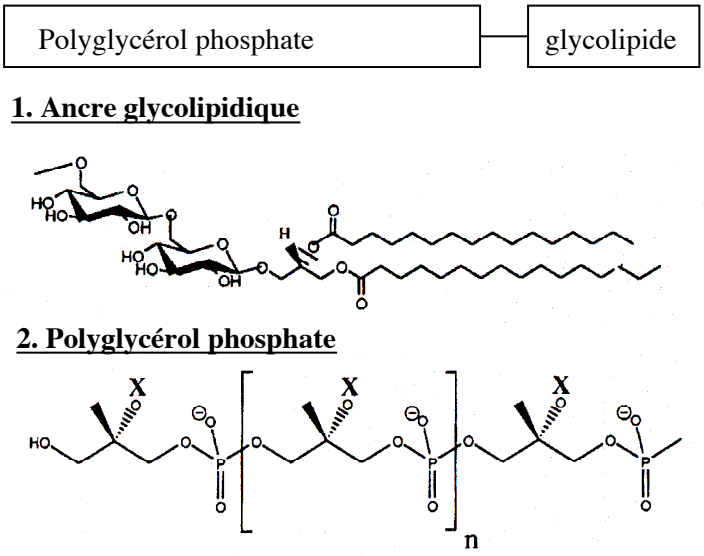


Figure 8 Structure d'un acide lipotéichoïque de type I (modifié de (Neuhaus and Baddiley, 2003)).

Chez les bactéries Gram+ à faible taux de G + C, la structure de l'acide lipotéichoïque la plus répandue consiste en une chaîne non branchée de polyglycérophosphates liés en 1-3, rattachée au glycolipide diacylglycérol par un lien phosphodiester (acide lipotéichoïque de type I). Les glycérophosphates sont partiellement substitués en position 2 par des esters de D-alanine et, éventuellement, par des résidus glycosilés (Fischer et al., 1997).

Les rôles des acides téichoïques et lipotéichoïques sont encore mal définis bien qu'ils fassent l'objet de nombreuses études.

Ils confèrent à la paroi la majeure partie de sa spécificité antigénique. En effet, les acides téichoïques constitueraient des sites de décorations spécifiques d'une espèce donnée, permettant à la bactérie de synthétiser une enveloppe distincte chimiquement de l'enveloppe d'autres microorganismes présentant un exosquelette de peptidoglycan identique.

De façon générale, les acides téichoïques et lipotéichoïques forment, avec le peptidoglycan, un réseau polyanionique contribuant aux propriétés d'élasticité, de porosité, de tension de surface et de répartition de charges de la paroi (Neuhaus and Baddiley, 2003) . Celle-ci est considérée comme un gel de polyélectrolytes possédant des propriétés d'échange d'ions nécessaires non seulement au maintien de l'homéostasie des cations métallique et à son contrôle mais aussi intervenant dans le trafic d'ions, de nutriments, de protéines et des antibiotiques. La paroi joue également un

rôle dans la liaison des protéines de paroi, dans la présentation des hydrolases de la muréine et des adhésines et détermine, en partie, l'hydrophobicité de surface de la cellule. En conclusion, l'enveloppe bactérienne peut être considérée comme un organelle fournissant les fonctions indispensables à la croissance de la bactérie Gram-+ dans sa niche écologique.

Chez les Gram+ à faible taux de G+C, les acides téichoïques (TA) et lipotéichoïques (LTA) sont souvent substitués par des esters de D-alanine protonés. En recherchant les rôles de ces esters de D-alanine, un certain nombre de fonctions peuvent être attribuées aux acides téichoïques et lipotéichoïques les portant (Neuhaus and Baddiley, 2003)

Les acides téichoïques substitués par des esters de D-alanine pourraient intervenir dans le maintien de l'homéostasie cationique et dans la capture d'ions métalliques nécessaires à des fonctions cellulaires. Dans diverses expériences, il existe une corrélation entre la proportion d'esters D-alanines et la capacité de liaison du magnésium dans la paroi. Plus il y a d'esters de D-alanine, moins il y a d'ions magnésium.

On leur attribue également un rôle dans la modulation de l'activité de certaines protéines de surface. En effet, une migration (ou transacylation) des esters à des localisations ou régions spécifiques de la paroi a été observée. Elle constituerait un mécanisme unique de transduction du signal pour moduler certaines activités protéiques, nécessitant un microenvironnement spécifique pour leur fonction. En effet, l'absence (ou présence) des esters D-alanine à des endroits spécifiques de la paroi pourrait constituer un mécanisme de ciblage des protéines régulées par une charge ionique locale, telles que les hydrolases de la muréine.

En fonction de l'espèce bactérienne considérée, les fonctions que nous venons de citer pourraient être limitées et des rôles supplémentaires pourraient être attribués à ces polymères anioniques, notamment dans le contexte de la pathogénie bactérienne (Poyart et al., 2003). En effet, les acides téichoiques sont considérés comme des inducteurs de certains médiateurs proinflammatoires. Outre le fait que ce sont des molécules immunogéniques, ils constituent également des activateurs du complément. Enfin, ils participeraient également à l'agrégation bactérienne et à la formation de biofilms via la présentation d'adhésines.

2.1.3. Les acides téichuroniques et les lipoglycans.

Moins fréquemment rencontrés que les acides téichoïques et lipotéichoïques, ces polymères anioniques sont fonctionnellement similaires (Neuhaus and Baddiley, 2003).

Les lipoglycans se distinguent des acides lipotéichoïques uniquement par l'absence de phosphate dans les sous-unités répétées. On les retrouve plus fréquemment chez les bactéries à haut taux de G+C (Fischer et al., 1997).

Chez *Micrococcus luteus*, le caractère anionique du lipomannan trouve son origine dans la présence de groupements succinyles estérifiés sur les résidus mannosyles.
En conditions de croissance particulières, on trouve des acides téichuroniques, chez des bactéries telles que *Bacillus subtilis*, *Bacillus licheniformis* ou *Micrococcus luteus*.
Chez *Bacillus subtilis*, la synthèse d'acide téichuronique est initiée lorsque la concentration en phosphate dans le milieu est faible. L'hypothèse fréquemment avancée pour expliquer ce changement de composition de la paroi est que la bactérie largue les acides téichoïques existants, libérant ainsi du phosphate dans le milieu. Celui-ci peut alors être utilisé pour des besoins urgents tels que la synthèse des acides nucléiques. Les acides téichuroniques de *Bacillus subtilis* sont des polymères d'acides glucuroniques et de n-acétyl-galactosamines (Robson and Baddiley, 1977). Ceux de *Micrococcus luteus* sont constitués d'acides N-acétyl-mannosaminuronique et de glucoses.
Même à faible concentration en phosphate, il faut noter qu'il y a toujours synthèse d'acide lipotéichoïque, ce dernier semblant donc indispensable pour la croissance cellulaire (Ward, 1981).

2.2. Les protéines de surface des bactéries Gram+

2.2.1. Présentation générale des protéines de surface des bactéries Gram+

Les protéines de surface des bactéries Gram+ sont présentées à l'extérieur pour se lier à un substrat localisé dans l'environnement immédiat. Ces protéines peuvent avoir des fonctions très diverses :

- croissance et *turnover* de la paroi cellulaire
- acquisition de nutriments
- liaison à des tissus hôtes
- liaison à des composants du système immunitaire
- maturation de protéines
- agrégation bactérienne en vue d'un transfert d'ADN par conjugaison

etc... (Navarre and Schneewind, 1999).

Les protéines sécrétées chez les Gram+ sont généralement marquées par un peptide-signal. Ce dernier se compose le plus souvent d'un coeur de résidus hydrophobes avec, à l'extrémité N-terminale quelques résidus chargés positivement. Ce peptide-signal est clivé par des peptidases spécifiques après translocation au travers de la membrane cytoplasmique. Pour la sécrétion, outre la présence de ce peptide-signal, la cellule peut faire appel à deux types de mécanismes.
Avant de décrire ces mécanismes, il est nécessaire de souligner que, contrairement aux études menées chez les bactéries Gram-, le processus par lequel les bactéries Gram+ sécrètent les protéines a été peu étudié. Le séquençage du génome de plusieurs bactéries Gram+ a montré que des gènes

de sécrétion, initialement identifiés chez *E. coli*, se retrouvent aussi chez les bactéries Gram+. On pense donc que les bactéries Gram+ possèdent des systèmes de translocation similaires à ceux étudiés chez *E. coli*. Ils sont actuellement au nombre de deux (Lee and Schneewind, 2001).
Dans un premier système, la translocation de la protéine d'intérêt a lieu après traduction complète de cette dernière (figure 9). Dans ce cas, elle se lie à une protéine chaperone de sécrétion (le plus souvent SecB, ou un équivalent non encore identifié chez *Bacillus subtilis* et d'autres bactéries Gram+), maintenant le polypeptide dans un état déplié. Une fois le mécanisme de sécrétion amorcé par le peptide-signal, SecB se dissocie, permettant la translocation de la protéine au travers de la membrane via un tunnel de translocation. Ce dernier est constitué des protéines SecYEG et l'énergie nécessaire à la translocation est fournie par la protéine SecA, une ATPase. Les protéines membranaires SecD, SecF et YajC sont associées au pore de translocation et semblent réguler la translocation SecA-dépendante.

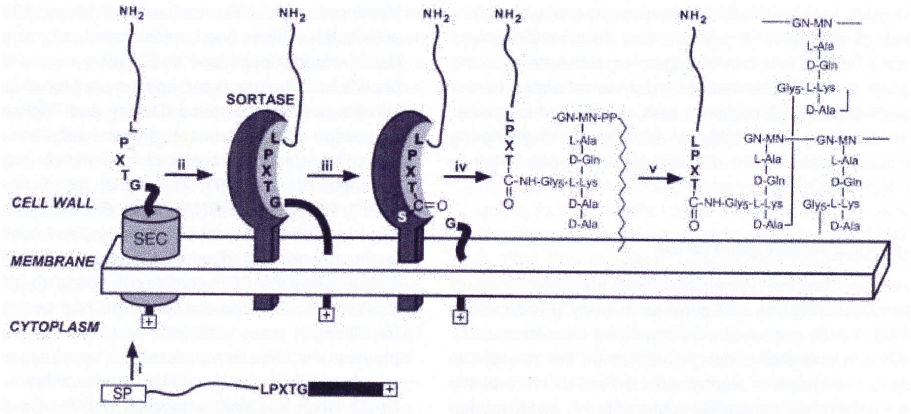


Figure 9 Illustration du système de translocation Sec (Ton-That et al., 2004).
(i) Dans la première étape, les protéines précurseurs portant un peptide-signal N-terminal sont transloquées au travers de la membrane cytoplasmique via le pore de translocation. (ii) Puis, le signal de sortie C-terminal freine la progression de la protéine, (iii) permettant de ce fait à la sortase de cliver le lien peptidique entre la thréonine et la glycine du motif LPXTG. Il y a ainsi formation d'un intermédiaire enzyme thioester. (iv) L'attaque nucléophile du groupement amine libre du lipide II sur la liaison thioester crée une liaison amide entre la protéine de surface et un pont de pentaglycine. (v) la protéine de surface liée au lipide est d'abord incorporée dans la paroi par une réaction de transglycosylation. La sous-unité composée du pentapeptide de glycine lié à la protéine de surface est ensuite lié aux autres peptides de la paroi , générant le tétrapeptide mature présent dans la muréine.

Dans un deuxième système, une particule de reconnaissance du signal (SRP) se lie au peptide-signal de la chaîne naissante et arrête momentanément la traduction de la protéine. Après liaison de cette protéine à son récepteur dans la membrane plasmique, la traduction de la protéine reprend, délivrant la

protéine en traduction au canal de translocation Sec. C'est probablement le mécanisme de traduction qui fournit la force ou l'énergie nécessaire à la translocation au travers de la membrane plasmique.
Après translocation, les protéines se replient, aidées par des protéines chaperones telles que PrsA (Sarvas et al., 2004). Cette protéine transmembranaire se retrouve chez toutes les bactéries Gram+.
Chez les bactéries Gram+, on a répertorié à ce jour sept peptidases du signal différentes, alors qu'il n'y en a que deux chez *Escherichia coli* (Comfort and Clubb, 2004).
Outre leur fonction, les protéines de surface des bactéries Gram+ peuvent être classées en fonction de la façon dont elles sont présentées en surface. Actuellement, on distingue cinq grandes classes qui seront brièvement présentées dans les paragraphes suivants (figure 10):

- les protéines liées de façon covalente à la paroi grâce à un motif LPXTG,
- les protéines insérées dans la membrane plasmique par une ancre hydophobe,
- les lipoprotéines liées de façon covalente à la membrane cytoplasmique,
- les protéines liées de façon non covalente par des liaisons électrostatiques,
- les protéines présentant des modules GW leur permettant d'interagir de façon non covalente avec l'acide lipotéichoïque de la paroi. Ces modules sont constitués d'unités répétées débutant par un dipeptide GW.

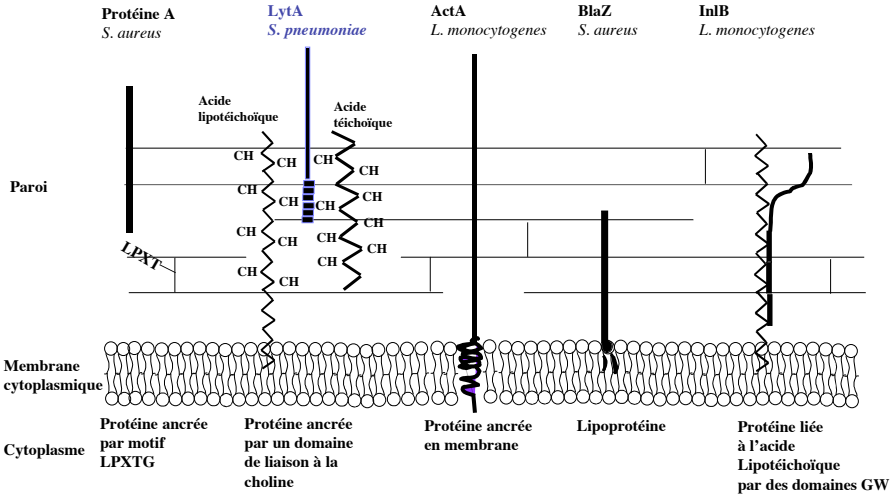


Figure 10 Principaux types de protéines de surface chez les bactéries à Gram+ (modifié de (Cossart and Jonquieres, 2000)).
La protéine A est liée de façon covalente à la paroi par un motif LPXTG. L'amidase LytA est attachée par des interactions ioniques aux résidus choline décorant les acides téichoïques et lipotéichoïques de *Streptococcus pneumoniae*. La protéine de polymérisation de l'actine ActA de *Listeria monocytogenes* est ancrée en membrane par une portion de séquence hydrophobe. La β-lactamase de *Staphylococcus aureus* (BlaZ) est liée de façon covalente à la membrane par une cystéine tandis que la protéine InlB s'associe de façon non covalente à l'acide lipotéichoïque par des domaines de liaison GW.

2.2.2. Les protéines liées de façon covalente au peptidoglycan par un motif LPXTG

Cette classe de protéines de surface présente un motif LPXTG en C-terminal, responsable de l'attachement covalent à la paroi cellulaire. Outre la séquence LPXTG, cette région, appelée "signal de sortie de la paroi", est composée d'un domaine hydrophobe (15 à 19 résidus) et d'une queue de cinq à neuf acides aminés principalement chargés (Sarvas et al., 2004).
Chez le staphylocoque doré, l'attachement de la protéine A se fait en quatre grandes étapes (figure 9) (Mazmanian et al., 2001).
Comme nous venons de le mentionner ci-dessus, la première étape consiste en la translocation à travers la membrane via le peptide-signal. Puis, la protéine semble freinée dans sa diffusion vers le milieu extracellulaire grâce à la présence du signal de sortie. Cette rétention permet la reconnaissance du motif LPXTG et son clivage entre la thréonine et la glycine par une sortase (trans-peptidase associée à la paroi). Enfin, la protéine libérée de son signal de sortie est liée de façon covalente à un précurseur de paroi lié au lipide II

qui est, à son tour, incorporé au peptidoglycan par des réactions de transpeptidation et transglycosilation (Comfort and Clubb, 2004).
Les protéines liées de façon covalente à la paroi présentent un arrangement structural commun comportant un ensemble de domaines répétés présentant ou non une activité, une série de prolines suivie du motif LPXTG (Navarre and Schneewind, 1999). Par contre, les domaines N-terminaux possédant des activités catalytiques ou de liaison sont très différents. Les fonctions de protéines de surfaces ancrées dans la paroi sont extrêmement diverses. Nous pouvons citer à titre d'exemple :
- la liaison à des immunoglobulines
- la liaison à des protéines du sérum ou de la matrice extracellulaire
- les enzymes participant, par exemple, à la dégradation en sous-unités plus petites de polymères nutritionnels larges et non transportables
- des protéines participant à l'agrégation bactérienne

2.2.3. Les protéines liées aux lipides de la membrane cytoplasmique

Ces protéines présentent à leur extrémité N-terminale un motif LX1X2C (où X1 est souvent une alanine, une sérine, une valine, une glutamine ou une thréonine et X2 est souvent une glycine ou une alanine). La protéine forme un lien covalent entre la cystéine du motif et le diacylglycérol de la bicouche lipidique de la membrane plasmique (Rigden et al., 2003).

2.2.4. Les protéines de surface possédant des modules GW

Ces protéines possèdent en C-terminal un certain nombre d'unités répétées débutant par un dipeptide GW. Les *repeats* forment des sites de liaison capables de se lier de façon non covalente à l'acide lipotéichoïque de certaines bactéries (Jonquières et al., 1999) (Braun et al., 1997).

2.2.5. Les protéines de surface ancrées dans la membrane plasmique

Ces protéines présentent une portion de séquence hydrophobe, suivies de résidus chargés, leur permettant de s'ancrer en membrane (Kocks et al., 1992).

2.2.6. Les protéines liées de façon non covalente par des interactions électrostatiques

La dernière grande famille des protéines de surface des bactéries Gram+ se lient à la paroi de façon non covalente. Ces protéines jouent notamment un rôle dans la synthèse et le *turnover* de la paroi (hydrolases de la muréine). Elles peuvent aussi constituer des facteurs de virulence (PspA de *Streptococcus pneumoniae*, InlB de *Listeria monocytogenes* comme invasine ou internaline) ou jouer le rôle de toxines (TcdA, TcdB de *Clostridium difficile*).
Elles ont comme caractéristique de présenter une structure en domaines bien distincts. En général, ces protéines arborent un peptide-signal N-terminal, suivi d'un domaine possédant l'activité catalytique. Ce dernier est flanqué en N-terminal ou C-terminal de structures répétées permettant le ciblage de la protéine vers son récepteur spécifique, qu'il soit sur la paroi bactérienne ou sur l'enveloppe d'une cellule cible.
Pour la compréhension de notre travail, nous nous pencherons plus en détail sur les hydrolases de la muréine puisque celles de *Streptococcus pneumoniae* consitueront le point de départ de ce travail.
Comme leur nom l'indique, les hydrolases de la muréine hydrolysent le peptidoglycan à des endroits particuliers. L'hydrolyse de la muréine peut avoir lieu pendant la croissance physiologique de la paroi. On parle alors d'autolysines puisque la bactérie synthétise une enzyme responsable de l'hydrolyse de sa propre paroi. Ces enzymes peuvent aussi servir à détruire la paroi d'espèces bactériennes mises en compétition et agissent alors comme bactériocines. C'est le cas de la lysostaphine de *Staphylococcus simulans* qui clive beaucoup plus activement le peptidoglycan de *Staphylococcus aureus* que son propre peptidoglycan (Baba and Schneewind, 1996). Enfin, l'hydrolase de la muréine LytA synthétisée par le phage 11 de *Staphylococcus aureus* permet le relarguage des particules phagiennes par hydrolyse de la paroi de la bactérie.
En fonction du type de lien clivé, on peut classer les hydrolases de la muréine en plusieurs catégories (Navarre and Schneewind, 1999) (figure 11) :

- les N-acétylmuramidases (muramidases ou lysozymes) clivent le lien glucosidique β (1-4) entre l'acide N-acétylmuramique et le N-acétylglucosamine,
- les N-acétylglucosamidases (glucosamidases) clivent le lien glucosidique β (1-4) entre le le N-acétylglucosamine et entre l'acide N-acétylmuramique,
- les N-acétylmuramoyl-L-alanine amidases (amidases) clivent le lien amide entre le groupement D-lactyl de l'acide muramique et le résidu L-alanine du pont peptidique,

- les endopeptidases sont responsables de l'hydrolyse des ponts peptidiques reliant les brins de glycan.

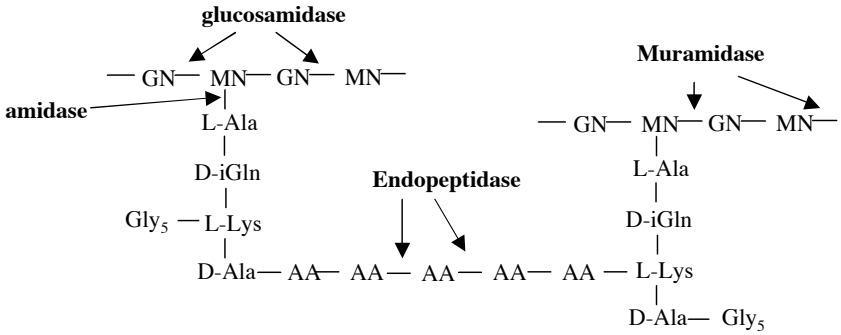


Figure 11 Illustration de l'activité enzymatique de diverses hydrolases de la muréine (modifié de (Navarre and Schneewind, 1999)).
GN = N-acétyl-glucosamine
MN = acide N-acétyl-muramique
AA = acide aminé

Certaines hydrolases de la muréine peuvent porter plusieurs activités enzymatiques. C'est le cas de l'hydrolase du phage g11 de *Staphylococcus aureus*. Elle clive les ponts peptidiques de la muréine soit entre l'acide muramique et la L-alanine (activité amidase) soit entre un D-alanine et la première glycine du pentapeptide (activité endopeptidase).
Les hydrolases de la muréine peuvent être synthétisées sous des formes très diverses :
- préproenzyme avec peptide-signal puis clivage de la proenzyme en domaines enzymatiques distincts et indépendants. C'est le cas de la lysostaphine de *Staphylococcus simulans* ou de l'autolysine Alt des staphylocoques
- proenzyme sans peptide-signal
- enzyme sans peptide-signal (amidase LytA de *Streptococcus pneumoniae*)
- protéine avec peptide-signal (PspA de *Streptococcus pneumoniae*).

2.3. La structure chimique de la paroi de *Streptococcus pneumoniae*

Epaisse et relativement homogène, la paroi de *Streptococcus pneumoniae* est typique des bactéries Gram+. Elle est constituée de couches de peptidoglycan sur lequel viennent se greffer des acides téichoïques et lipotéichoïques.
La structure de la muréine est classique et consiste en la polymérisation d'un dissacharide NAG-NAM portant un pentapeptide de base (Fischer and Tomasz, 1985). Il existe cependant une particularité : 84% des N-acétyl-glucosamine et 10% des acides N-acétyl-muramiques sont déacétylés.

L'enzyme responsable de cette déacétylation, PgdA, a été récemment identifiée (Vollmer and Tomasz, 2001). La liaison entre les brins de glucan se ferait par un pont peptidique reliant les peptides de base de deux brins de glucan adjacents. Ce pont serait composé d'alanine, sérine, glycine, acide aspartique et présenterait une variabilité dans sa composition (Fischer and Tomasz, 1985).

Par contre, la structure chimique des acides téichoïques et lipotéichoïques de *Streptococcus pneumoniae* se distingue très nettement de celle d'autres bactéries Gram+, et ce, pour diverses raisons :

- d'abord, c'est une des seules Gram+ connues dont la structure des acides téichoïques et lipotéichoïques est identique (Fischer et al., 1993),
- ensuite, le glycérophosphate, habituellement retrouvé dans l'acide lipotéichoïques des Gram+ à faible taux de G+C, est remplacé par du ribitolphosphate (figure 12).

Entre ces unités de ribitol, est intercalé un tétrasaccharide contenant du D-glucose, du 2-acétamido-4-amino-2,4,6-tridéoxy-D-galactose (AATGal) chargé positivement et de deux molécules de N-acétyl-galactosamine portant chacun un ou deux résidus de phosphocholine en fonction de la souche considérée. Les unités répétées de ribitolphosphate-tétrasaccharide, en moyenne six par brin d'acide lipotéichoïque, sont liées par des liens phosphodiesters entre le O_5 du ribitol et le O_6 du résidu glucose de l'unité adjacente (Fischer et al., 1997) (Behr et al., 1992).

La chaîne hydrophile de l'acide lipotéichoïque, constituée en moyenne de 5 à 8 unités répétées, est liée par un lien phosphodiester à un unique glycolipide de diacylglycérol par un trisaccharide composé de deux résidus D-glucopyranosyles entre lesquels est inséré un résidu de 2-acétamido-4-amino-2,4,6-tridéoxy-D-galactose (AATGal). Cette ancre lipidique n'a jamais été trouvée à l'état libre parmi les glycolipides membranaires identifiés du pneumocoque.

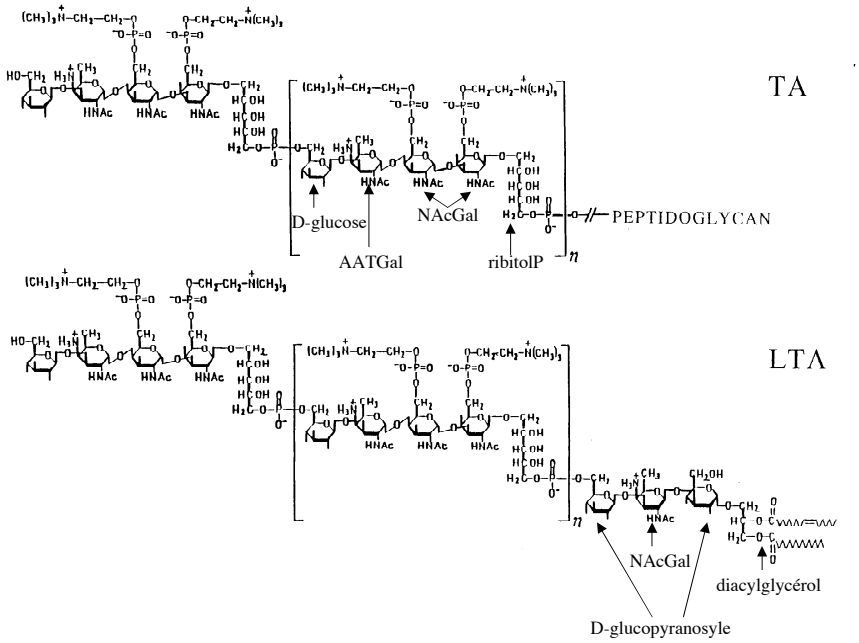


Figure 12 Structure de l'acide téichoïque (TA) et de l'acide lipotéichoïque (LTA) de la souche R6 de *Streptococcus pneumoniae* (modifié de (Tomasz and Werner, 2000)).
TA : acide téichoïque, LTA : acide lipotéichoïque, AATGal : 2-acétamido-4-amino-2,4,6-tridéoxy-D-galactose, NacGal : N-acétyl-galactosamine.

2.4. La choline pariétale chez *Streptococcus pneumoniae*

La choline est un facteur de croissance indispensable chez *Streptococcus pneumoniae*, la bactérie étant incapable de la synthétiser *de novo*.
Actuellement, six espèces bactériennes possédant de la choline dans leur paroi ont été répertoriées (tableau 2) (Garcia et al., 1998).

Bactéries	Références
Streptococcus pneumoniae	(Tomasz, 1967)
Streptococcus oralis	(Ronda et al., 1991)
Streptococcus constellatus	(Kilpper-Bätz et al., 1985)
Streptococcus mitis	(Gillespie et al., 1993)
Clostridium beijerinckii	(Garcia et al., 1988)1988
Clostridium NI-4	(Garcia et al., 1988) (Podvin et al., 1988)

Tableau 2 Liste des bactéries possédant de la choline dans leur paroi (Garcia et al., 1998).

Les fonctions de la choline dans la paroi de *Streptococcus pneumoniae* font l'objet d'un certain nombre d'investigations, particulièrement en ce qui concerne les conséquences biologiques de sa substitution par de l'éthanolamine.
En remplaçant la choline dans le milieu de culture par un analogue tel que l'éthanolamine, Briese et ses collaborateurs ont observé la formation de longues chaînes de bactéries suite à la non séparation des cellules-filles après multiplication et une perte de la capacité de transformation des cellules (Briese and Hakenbeck, 1985). Ils ont également constaté qu'il n'y avait plus de lyse spontanée en fin de phase stationnaire ni de lyse induite par la pénicilline ou par le déoxycholate. Comme nous le verrons par la suite, certains de ces phénotypes sont à mettre en rapport avec l'inactivation et/ou le non-ancrage d' hydrolases de la muréine, normalement présentes à la surface des bactéries. Les phosphorylcholines sont donc nécessaires pour l'ancrage de toute une série de protéines de surface possédant un domaine de liaison à la choline : les " *choline-binding proteins* " ou CBP.
Il semble que la choline joue aussi un rôle dans la pathogénie du pneumocoque. En effet, *Streptococcus pneumoniae* est capable de variations de phase spontanées et réversibles, marquées par la formation de colonies transparentes ou opaques (Weiser et al., 1994). Ces variations semblent liées à la capacité du pneumocoque à envahir certains types de tissus. Dans ce contexte, il semble que les pneumocoques formant des colonies transparentes possèdent plus d'acides téichoïques et plus de choline dans leur paroi que les colonies opaques (Kim and Weiser, 1998).
Toujours concernant la pathogénie de *Streptococcus pneumoniae*, les phosphorylcholines de la paroi semblent interagir directement avec la surface des cellules hôtes et joueraient donc un rôle dans l'attachement et l'invasion (Tuomanen, 1999). Des résultats obtenus par Cundell viennent renforcer cette hypothèse (Cundell et al., 1995). En effet, les phosphorylcholines serviraient de ligand adhésif aux récepteurs du « Platelet-Activating Factor » présents à la surface de diverses cellules épithéliales et endothéliales. La liaison des phosphorylcholines à ces récepteurs augmenterait l'adhésion du pneumocoque et donc sa capacité d'envahir les cellules hôtes.
Enfin, les acides téichoïques contenant de la choline sont essentiels pour l'adsorption du phage Dp-1 (Lopez et al., 1982).
Les mécanismes impliqués dans l'incorporation de choline dans la paroi ne sont pas encore bien compris. En 1996, Whiting et Gillespie ont proposé une voie métabolique hypothétique, par analogie avec des voies métaboliques impliquant des cholines kinases chez *Escherichia coli* et *Saccharomyces cerevisiae* (Whiting and Gillespie, 1996). Cette voie métabolique comprendrait trois étapes distinctes (figure 13) :
1) après transport de la choline dans la cellule, il y aurait phosphorylation de la choline par une choline kinase,
2) la choline-phosphate serait activée par une choline cytidylyltransférase en

CDP-choline,

3) la partie choline-phosphate serait transférée sur les unités répétées de tétrasaccharide-ribitol phosphate des acides téichoïques et lipotéichoïques naissants par une choline-phosphate-phospho-transférase. Ces auteurs ont mis en évidence une activité choline kinase chez *Streptococcus pneumoniae*.

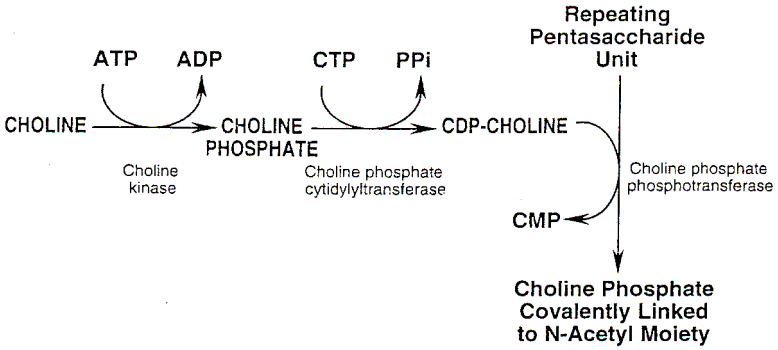


Figure 13 Voie d'incorporation hypothétique de la choline dans la paroi de *Streptococcus pneumoniae* (Whiting and Gillespie, 1996).

Par la suite, la découverte de phosphorylcholines dans le LPS d'*Haemophilus influenzae* et la mise en évidence du locus *lic1* impliqué dans l'expression des épitopes de phosphorylcholines chez cette bactérie ont permis, par analogie, d'approfondir les connaissances sur le mécanisme d'incorporation de la choline dans la paroi du pneumocoque.

Chez *Haemophilus influenzae*, des variations de phase spontanées et très fréquentes des épitopes de phosphorylcholines permettent à la bactérie de persister dans le tractus respiratoire et de résister à des protéases du sérum (Weiser et al., 1997). L'expression de ces épitopes dépend d'un locus *lic1* composé de quatre gènes.

LicA et *licB* codent respectivement pour une choline kinase et un transporteur de choline (Weiser et al., 1997). LicC est homologue à des nucléotides pyrophosphorylases. LicD est homologue à une protéine de pneumocoque, codée par le gène *cps*G situé dans un locus responsable de la biosynthèse de polysaccharides capsulaires. La fonction exacte de CpsG est inconnue.

Chez *Streptococcus pneumoniae*, on a mis en évidence un locus *lic* contenant des gènes homologues à ceux du locus *lic1* (Zhang et al., 1999) (figure14). Orientés dans le sens inverse des gènes *licA, licB et licC*, les gènes *licD1* et *licD2* sont séparés de ceux-ci par trois phases ouvertes de lecture.

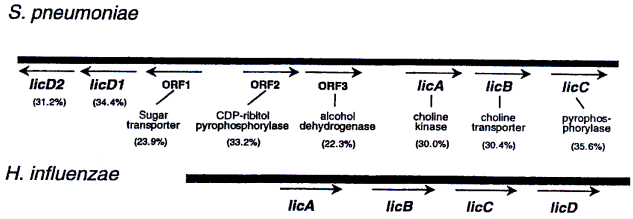


Figure 14 Comparaison de l'organisation des loci *lic* chez *Streptococcus pneumoniae* et *Haemophilus influenzae*. La direction des phases ouvertes de lecture est indiquée par des flèches. Les fonctions potentielles des produits des gènes, basées sur la similarité de séquence, sont signalées (Zhang et al., 1999).

Par comparaison de séquences, on a assigné aux protéines LicA, LicB et LicC de *Streptococcus pneumoniae* des fonctions similaires à celles attribuées aux protéines *d'Haemophilus influenzae*. Les gènes *licD1* et *licD2* sont homologues au gène *cpsG* du pneumocoque. La délétion du gène *licD2* induit des changements phénotypiques assez importants dans la capacité de transformation de la souche, la lyse induite par la pénicilline, la lyse en phase stationnaire, la virulence et l'assimilation de choline. Une analyse de la composition en choline des acides lipotéichoïques de ce disruptant montre qu'il contient la moitié moins de choline qu'une souche sauvage et que la souche incorpore deux fois moins de choline qu'une souche sauvage. Cette observation suggère que le gène *LicD2* coderait pour une protéine de transfert de la choline au départ d'un donneur (CDP-choline ou CDP-lécithine) vers l'acide lipotéichoïque et, peut-être, vers l'acide téichoïque. La protéine serait donc responsable de l'incorporation d'une des deux molécules de choline habituellement présentes sur les unités répétées composant les acides téichoïques et lipotéichoïques du pneumocoque. Bien qu'aucun mutant *LicD1* n'ai pu être isolé, Zhang et ses collaborateurs suggèrent un rôle d'incorporation de la deuxième molécule de choline par la protéine LicD1 (Zhang et al., 1999). L'existence de deux protéines remplissant la même fonction d'incorporation de choline dans la paroi du pneumocoque suggère l'existence de deux classes de choline fonctionnellement distinctes. Dans ce contexte, une classe fonctionnelle de phosphorylcholines, liée à la fonction de licD1, servirait d'ancrage pour le CBP tandis que la seconde classe dépendant de l'activité de licD2 servirait à activer l'autolysine LytA de *Streptococcus pneumoniae*.

Il est intéressant de noter que, dans des expériences de dénaturation thermique visant à étudier l'organisation structurale de l'amidase LytA, Usobiaga et ses collaborateurs ont observé qu'à partir de 10 mM de choline, la dénaturation de LytA est biphasique (Usobiaga et al., 1996).

Selon les auteurs, ce profil de dénaturation peut s'expliquer si on suppose l'existence de trois domaines de coopération sur les deux domaines de liaison du dimère :

- deux domaines indépendants et identiques C1; situés sur les bords du

dimère et se dénaturant à basse température,
- un bloc dimérique de coopération $(C2)_2$ se dénaturant à plus haute température.

Ces résultats amènent les auteurs à supposer l'existence de deux classes de sites de liaison se distinguant par leur affinité pour la choline. D'autres expériences de dichroïsme circulaire et de Calorimétrie différentielle à balayage ("Differential Scanning Calorimetry" ou DSC), réalisées sur l'amidase LytA ou sur son domaine de liaison ClytA en présence ou absence de choline, mettent également en évidence l'existence de ces deux types de sites (Medrano et al., 1996). La caractérisation structurale et thermodynamique de l'amidase Pal du phage Dp-1 semble confirmer leur existence bien que rien de tel ne soit observé pour l'amidase EJ1 (Varea et al., 2004) (Saiz et al., 2002).

Ces expériences pourraient renforcer l'hypothèse de deux classes de molécules de choline fonctionnellement distinctes.

Comme nous l'avons signalé au début de ce paragraphe, la choline est un facteur de croissance indispensable chez *Streptococcus pneumoniae*. Une hypothèse avancée pour expliquer cette observation consiste à relier la présence de choline (ou d'un analogue structural tel que l'éthanolamine) à la synthèse du peptidoglycan. Des expériences ont montré que la synthèse du peptidoglycan s'arrête lorsqu'il n'y a plus de choline et qu'elle reprend lorsqu'on ajoute de la choline ou de l'éthanolamine dans le milieu (Fischer and Tomasz, 1985). Or, pour être transloqué au travers de la membrane plasmique, les acides téichoïques et les dissacharides NAG-NAM-pentapeptide utilisent le même transporteur lipidique : l'undécaprénol-phosphate. On pourrait donc penser que les acides téichoïques ne possédant pas de choline (ou d'éthanolamine) ne sont pas transférés vers la paroi et qu'ils s'accumulent sous forme liées à l'undécaprénol-phosphate, faisant de ce fait diminuer le stock de ce transporteur lipidique. En conséquence, il y aurait arrêt de la synthèse du peptidoglycan et mort cellulaire (Fischer, 2000).

En conclusion, le métabolisme de la phosphocholine joue un rôle significatif dans le métabolisme et la pathogénie de *Streptococcus pneumoniae*.

2.5. Les protéines de surface possédant un domaine de liaison à la choline

Les principales protéines se liant à la choline ont été étudiées chez *Streptococcus pneumoniae*, chez les phages Dp-1, Cp-1, Cp-9, EJ-1 et HB3 du pneumocoque et chez *Clostridium Beijerinckii*. Le tableau 3 reprend les protéines de liaison à la choline actuellement connues. Pour plus de clarté, nous ne détaillerons, par la suite, que des données relatives aux protéines de liaison à la choline de *Streptococcus pneumoniae*.

Protéine	fonction	nombre de motifs répétés du domaine de liaison à la choline	espèce bactérienne ou espèce phagienne
CbpA	adhésine	9 motifs	*Streptococcus pneumoniae*
CPL1	hydrolase de la muréine	6 motifs	phage Cp-1
CPL9	hydrolase de la muréine	6 motifs	phage Cp-9
CspA	hydrolase de la muréine	4 - 5 motifs	*Clostridium beijerinckii*
CspB	fonction inconnue	6 motifs	*Clostridium beijerinckii*
CspC	fonction inconnue	4 motifs	*Clostridium beijerinckii*
CspD	fonction inconnue	5 motifs	*Clostridium beijerinckii*
EJL	hydrolase de la muréine	6 motifs	phage EJ-1
HBL3	hydrolase de la muréine	6 motifs	phage HB-3
LytA	hydrolase de la muréine	6 motifs	*Streptococcus pneumoniae*
LytB	hydrolase de la muréine	15 motifs	*Streptococcus pneumoniae*
LytC	hydrolase de la muréine	11 motifs	*Streptococcus pneumoniae*
PbcA	C3-binding protein	4 motifs	*Streptococcus pneumoniae*
PcpA	fonction inconnue	12 motifs	*Streptococcus pneumoniae*
PcpC	fonction inconnue	16 motifs	*Streptococcus pneumoniae*
PAL	hydrolase de la muréine	5 motifs	phage Dp-1
PspA	facteur de virulence	10 motifs	*Streptococcus pneumoniae*
PspC	fonction inconnue	11 motifs	*Streptococcus pneumoniae*
SpsA	se lie à IgA sécrétoire	4 motifs	*Streptococcus pneumoniae*
Pce = CbpE	phosphrylcholine estérase	10 motifs	Streptococcus pneumoniae
	lysine du phage SM1	5 motifs	phage *SM1*
SpaA	antigène de surface protecteur	7 motifs	*Erysipelotrhix rhusiopathiae*
CbpC	fonction inconnue	9 motifs	*Streptococcus pneumoniae*
CbpD	fonction inconnue	5 motifs	*Streptococcus pneumoniae*
CbpF	fonction inconnue	6 motifs	*Streptococcus pneumoniae*
CbpG	fonction inconnue	3 motifs	*Streptococcus pneumoniae*
CbpI	fonction inconnue	6 motifs	*Streptococcus pneumoniae*
CbpJ	fonction inconnue	8 motifs	*Streptococcus pneumoniae*
pneumococcal surface protein, putative	fonction inconnue	9 motifs	*Streptococcus pneumoniae*
hypothetical protein	fonction inconnue	9 motifs	*Streptococcus pneumoniae*
hypothetical protein	fonction inconnue	11 motifs	*Clostridium perfringens*
unknown protein	fonction inconnue	10 motifs	*Clostridium perfringens*
PspA	fonction inconnue	4 motifs	*Clostridium perfringens*

Tableau 3 Liste des *choline-binding proteins* actuellement répertoriées.

2.5.1. Les 'choline-binding proteins'(CBP)

Le séquençage du génome de *Streptococcus pneumoniae* a été réalisé sur la souche encapsulée TIGR4 (Tettelin et al., 2001) et sur les souches non encapsulées R6 (Hoskins et al., 2001) et G54 (Dopazo et al., 2001). Il a permis d'identifier de nombreux gènes codant pour des protéines se liant à la choline. En fonction de la souche étudiée, on en dénombre douze (Hoskins et al., 2001) ou quinze (Tettelin et al., 2001). Les gènes sont répartis aléatoirement sur l'ensemble du chromosome du pneumocoque et ne présentent pas de séquence consensus au niveau du promoteur.
Parmi ces CBP, on retrouve des hydrolases de la muréine et des facteurs de virulence.
Les hydrolases de la muréine actuellement répertoriées chez *Streptococcus pneumoniae* sont :

- l'autolysine majeure LytA possédant une activité N-acétylmuramoyl-L-alanine amidase (amidase),
- LytB à activité N-acétylglucosamidase (glucosamidase),
- LytC présentant une activité de type lysozyme.

Outre l'enzyme Pce (ou phosphorylcholine estérase) dont l'action se situe aussi au niveau de la paroi, le séquençage du génome n'a pas permis de

mettre en évidence l'existence d'hydrolases de la muréine présentant une redondance de fonction. Ce phénomène est en contradiction avec les observations dans d'autres systèmes bactériens où plusieurs hydrolases de la muréine présentent la même spécificité enzymatique. Bien que le génome du pneumocoque ait été séquencé, il ne faut cependant pas exclure l'existence d'autres hydrolases de la muréine dépendantes ou indépendantes de la choline (De Las Rivas et al., 2002).

Les hydrolases de la muréine LytA, LytB et LytC interviennent dans des fonctions physiologiques cellulaires associées à la croissance de la paroi, son *turn-over* et la séparation des cellules filles en fin de division cellulaire. La fonction majeure de ces enzymes, la dégradation de la paroi, a des conséquences physiologiques importantes telles que la lyse cellulaire en fin de phase stationnaire, conduisant directement à la mort de la cellule (Jedrzejas, 2001). Comme nous le verrons par la suite, on attribue aussi aux hydrolases de la muréine des rôles dans la virulence (Di Guilmi and Dessen, 2002).

D'autres protéines de liaison à la choline, ne possédant pas d'activité hydrolase de la muréine, sont considérés uniquement comme des facteurs de virulence. C'est notamment le cas de CbpA et PspA. CbpA ou « choline-binding protein A » est encore appelée dans la littérature PspC (Brooks-Walter et al., 1999), SpsA (Hammerschmidt et al., 2000) et PbcA (Cheng et al., 2000). C'est la plus abondante des CBP du pneumocoque. On lui attribue diverses fonctions dont un rôle dans la colonisation du nasopharynx en tant qu'adhésine (Rosenow et al., 1997), la liaison à la portion C3 du complément (Cheng et al., 2000) et la liaison à des IgA sécrétoires (Hammerschmidt et al., 2000). PspA (ou Pneumococcal Surface Protein A) semble jouer un rôle de protection contre le système du complément de l'hôte (Jedrzejas, 2001) (Ren et al., 2003) et pourrait stabiliser la capsule du pneumocoque (Briles et al., 1998).

Peu de choses sont connues sur la régulation de l'expression des "*choline-binding proteins*" en général.

Ogunniyi et ses collaborateurs ont montré par la technique de RT-PCR que les facteurs de virulence PspA et Cbpa sont régulés indépendamment *in vivo*, particulièrement pendant les premiers stades de l'infection (Ogunniyi et al., 2002). Le système de transduction du signal à deux composants RR06/HK06 vient d'être mis en évidence pour la régulation de l'expression de CbpA (Standish et al., 2005).

Enfin, on pense que certaines protéines de liaison à la choline contribuent aux propriétés de surface de la bactérie, en termes d'hydrophobicité et de répartition de charges.

L'organisation générale des protéines se liant à la choline comprend deux aspects particuliers.

A l'exception des protéines PspA, LytB et LytC, une première caractéristique est l'absence de peptide-signal N-terminal permettant

l'exportation de ces protéines vers la paroi. On ne sait donc pas comment les CBP sont transloquées au travers de la membrane cytoplasmique.
Une deuxième grande caractéristique des protéines se liant à la choline est leur organisation modulaire, suggérant la fusion de deux éléments génétiques distincts lors de l'évolution. En effet, des analyses génétiques et biochimiques ont démontré que ces enzymes sont formées de deux modules fonctionnels distincts :

- un module catalytique, le plus souvent N-terminal, qui contient le site actif de l'enzyme
- un module de liaison, le plus souvent C-terminal, responsable de la reconnaissance du substrat c'est-à-dire dans le cas des hydrolases de la muréine, de la reconnaissance des phosphorylcholines portées par les acides téichoïques et lipotéichoïques de la paroi du pneumocoque (Sanz et al., 1996).

Quand les modules sont isolés l'un de l'autre, ils conservent leur capacité catalytique ou de liaison (Garcia et al., 1990) (Sanz et al., 1992). La construction d'enzymes chimériques portant le domaine catalytique d'une hydrolase de la muréine et le domaine de liaison d'une autre hydrolase a permis de confirmer cette caractéristique (Lopez et al., 1995) (Sanz et al., 1996) (Diaz et al., 1990).

2.5.2. La structure des domaines de liaison à la choline

Les domaines de liaison à la choline sont caractérisés par la présence de séquences répétées d'une vingtaine d'acides aminés (encore appelés *repeats* ou motifs répétés). La taille d'un domaine est très variable d'une protéine à l'autre puisque, parmi les protéines actuellement répertoriées, elle va de 4 à 16 *repeats*. Les domaines de liaison composés d'unités répétées constituent une caractéristique commune de nombreuses protéines impliquées dans des processus de reconnaissance. Ces dernières années, les techniques moléculaires ont permis d'identifier des *repeats* dans un grand nombre de protéines se liant à des ligands, surtout chez les Gram+ (Wren, 1991). Nous pouvons par exemple citer la famille des protéines M antiphagocytaires de *Streptococcus pyogenes* ou encore une famille de protéines se liant aux immunoglobulines chez les streptocoques, sans oublier les hydrolases de la muréine chez d'autres espèces de bactéries Gram+ que *Streptococcus pneumoniae*.
Les domaines de liaison constitués de séquences répétées ont été classés en différentes familles et leurs séquences ont été alignées afin de dégager des séquences consensus-clefs. Cette approche a été réalisée par deux auteurs différents, pour les domaines de liaison à la choline des protéines LytA et PspA de *Streptococcus pneumoniae* ainsi que pour les domaines de liaison à la choline des phages Cp-1, Cp-9 et HB-3 du pneumocoque.

Dans un premier type d'alignement, Wren a aligné les séquences répétées des toxines A et B *de Clostridium difficile*, celles d'une protéine se liant au glucan (GBP) *de Streptococcus mutans*, deux glucosyltransférases (GTF) de *Streptococcus downei* et *Streptococcus mutans*, l'amidase LytA de *Streptococcus pneumoniae* et les hydrolases de la muréine de phages HB-3, Cp-1 et Cp-9 du pneumocoque (Wren, 1991). Outre le fait que ces protéines possèdent des séquences répétées, elles se lient toutes à de petites molécules hydrophiles (choline ou saccharide) ou à des polymères tels que le glucan. De plus, on retrouve des réactions immunologiques croisées entre certaines de ces protéines. Par exemple, un antisérum dirigé contre la GBP de *Streptococcus mutans* cross-réagit avec la toxine A de *Clostridium difficile*. Le consensus établi après alignements des *repeats* de ces protéines est le suivant :

KAVTGWxTIxGxxYYFxxNGx

Cette étude permet de dégager les caractéristiques suivantes :

- conservation du dipeptide tyrosine-phénylalanine (YF) et d'une glycine (G) dix résidus en amont de ce dipeptide,
- conservation d'aromatiques à d'autres positions (surtout les deux positions en amont du dipeptide),
- alternance de résidus chargés et polaires avec des résidus promoteurs de *turn*.

Dans une autre expérience, Giffard et Jacques ont aligné les domaines de liaison de six glucosyltransférases des streptocoques, d'une protéine de liaison au glucan de *Streptococcus mutans*, de la toxine A de *Clostridium difficile*, de l'amidase LytA et de la protéine de surface PspA de *Streptococcus pneumoniae* (Giffard and Jacques, 1994). La taille des séquences répétées à aligner n'étant pas constante, ils définissent des régions caractérisées par une classe d'acides aminés plutôt qu'un consensus position par position. Le résultat obtenu est le suivant :

région 1 : 4 à 6 résidus incluant généralement une glycine, un aspartate ou une asparagine,
région 2 : cluster de 1 à 4 résidus incluant le plus souvent des tyrosines,
région 3 : 3 à 4 résidus très peu conservés,
région 4 : une glycine,
région 5 : 2 résidus peu conservés,
région 6 : un résidu hydrophobe,
région 7 : 1 résidu peu conservé,
région 8 : un résidu polaire neutre (souvent une glycine),
région 9 : un résidu peu conservé,
région 10 : 3 résidus incluant 1 à 3 acides aminés hydrophobes.

De nouveau, on constate l'existence d'un groupe relativement conservé de résidus aromatiques (région 2) ainsi que la présence de glycines à des positions précises (régions 4 et 8).
Bien que ces alignements n'incluent que très peu de protéines se liant à la choline, ils nous apprennent que les séquences répétées constituant un domaine de liaison peuvent être caractérisées par une séquence consensus après alignement. Ce consensus permet parfois de définir des résidus particulièrement bien conservés qui pourraient être, de ce fait, fonctionnellement ou structuralement importants.

2.5.3. L'amidase LytA de Streptococcus pneumoniae

Après cette introduction générale aux protéines se liant à la choline chez *Streptococcus pneumoniae*, nous consacrerons le chapitre suivant à décrire plus en détail les données déjà connues sur l'autolysine majeure LytA du pneumocoque.
En effet, c'est au départ du domaine de liaison à la choline de cette protéine que nous allons tenter d'élaborer un *tag* de purification.
Le choix de LytA comme protéine modèle repose sur le fait que c'est l'hydrolase de la muréine la plus étudiée. Les données expérimentales la concernant sont donc relativement nombreuses. Sa structure, inconnue lors de l'initiation de ce travail, a été publiée lorsque nous terminions notre analyse bioinformatique (*).
De par sa fonction d'hydrolase de la muréine, l'autolysine majeure joue un rôle dans l'élargissement et le *turn-over* de la paroi. Le gène codant pour cette protéine n'est cependant pas essentiel puisqu'un délétant Δ*LytA* présente un taux de croissance normal (Sanchez-Puelles et al., 1986). Il n'y a donc pas de fonction physiologique essentielle affectée. Par contre, un certain nombre d'autres phénotypes sont associés à ce délétant. D'abord, on constate la formation de petites chaînes composées de 6 à 8 cellules au lieu des diplocoques classiquement observés. Un rôle de séparation des cellules-filles en fin de division cellulaire a été attribué à cette enzyme. Ce rôle serait cependant secondaire puisqu'il a été mis en évidence que cette activité est surtout dévolue à la glucosamidase LytB (De Las Rivas et al., 2002). Le délétant Δ*LytA* ne subit plus de lyse en fin de phase stationnaire, ce qui pourrait constituer un désavantage puisqu'il n'y a plus libération d'ADN pouvant servir de source dans un phénomène de transformation naturelle. De même, le délétant développe une tolérance vis-à-vis des antibiotiques β-lactames. Bien que la fonction exacte de l'amidase dans l'effet bactériolytique de ces antibiotiques n'ait pas été expliquée, son rôle est confirmé. Les différentes fonctions que nous venons de citer ont été avalisées par des expériences de complémentation. (Ronda et al., 1987).
La contribution de l'autolysine majeure à la virulence de *Streptococcus pneumoniae* est également très étudiée mais son rôle précis est toujours en

(*) pour des raisons de clarté, la structure du domaine ClytA sera présentéé à la fin du chapitre consacré aux analyses bioinformatiques (Résultats - partie I).

discussion. L'introduction de mutations dans le gène *lytA* diminue significativement la virulence de la bactérie par rapport à une souche sauvage dans une expérience de compétition chez la souris (Berry and Paton, 2000). D'autres études montrent que l'autolysine LytA est capable d'induire une réponse protectrice chez la souris quand on infecte les poumons de cette dernière avec du pneumocoque (Berry et al., 1992). De façon plus indirecte, l'autolysine contribuerait au relarguage de toxines cytoplasmiques létales dont la plus connue est la pneumolysine (Balachandran et al., 2002). De nouveau, en fonction de la souche étudiée, ce rôle reste controversé.
Enfin, l'activité autolytique de LytA sur la paroi permet le relarguage de fragments pariétaux très inflammatoires pouvant contribuer à la pathogénie (Berry et al., 1989) (Jedrzejas, 2001). En effet, ces fragments pariétaux, et plus spécialement des produits de dégradation du peptidoglycan, peuvent être considérés comme des motifs moléculaires associés aux pathogènes, spécialement reconnus par des récepteurs, appelés de façon générique des « *Pathogen Recognition Receptors* ou PPR ». Le dipeptide D-glutamate – acide méso-diaminopimélique activerait spécifiquement le récepteur NOD1 tandis que le fragment acide muramique-L-alanine-D-glutamate-L-Lysine activerait le récepteur NOD2. De manière générale, l'activation des PPR, dont font partie NOD1 et NOD2, peut conduire à la mobilisation de molécules solubles de défense, à la mort des cellules ou tissus infectés, à l'induction de molécules co-stimulées par des cellules présentatrices d'antigènes ou à d'autres réponses physiologiques (Martinon and Tschopp, 2005) (Kufer et al., 2005).

Avant de passer en revue les expériences qui ont permis de mieux caractériser d'un point de vue structural le domaine de liaison à la choline, nous devons mentionner une particularité concernant l'activité enzymatique de l'amidase du pneumocoque. LytA peut exister sous forme active (forme C) ou sous forme inactive (forme E) (Holtje and Tomasz, 1975). En effet, si l'on cultive *Streptococcus pneumoniae* sur un milieu contenant de l'éthanolamine au lieu de la choline, les cellules contiennent la forme inactive de l'amidase et les parois contenant l'éthanolamine sont résistantes à son activité. La forme inactive, qui peut être aussi obtenue par synthèse de l'enzyme chez *Escherichia coli*, peut être convertie en forme active par incubation à 4°C de l'enzyme avec son substrat naturel (parois contenant de la choline) ou avec de la choline 2%. Ce phénomène, appelé la « conversion », est caractéristique de certaines hydrolases de la muréine telles que LytA ou Pal mais pas de toutes (Sheehan et al., 1997). La liaison de l'enzyme à son substrat naturel est donc un prérequis pour observer son activité catalytique. La choline est considérée comme un ligand allostérique qui, après liaison, induit des changements structuraux au niveau du domaine catalytique, rendant de ce fait l'enzyme active.

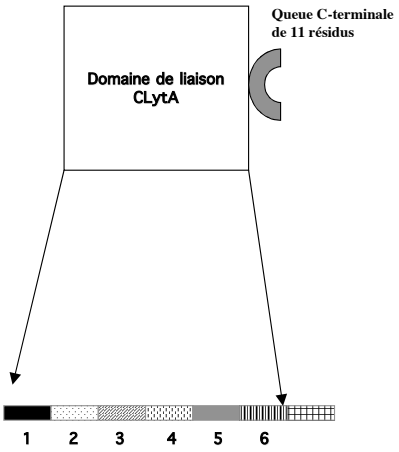


Figure 15 Représentation schématique du domaine de liaison ClytA, composé de six séquences répétées et d'une queue C-terminale.

De nombreuses études ont également été réalisées sur le mode d'attachement, relativement inhabituel parmi les bactéries Gram+, de l'autolysine à la paroi. Ces études portent donc plus spécifiquement sur le domaine de liaison à la choline. Rappelons que ce domaine est constitué de six séquences répétées et d'une queue C-terminale composée de onze résidus (figure 15). Outre des expériences de délétions progressives sur ce domaine, l'état oligomérique et la forme de la protéine entière et de son domaine C-terminal ont été caractérisés par ultracentrifugation analytique (Garcia et al., 1994) (Varea et al., 2000). Un examen de l'influence de la choline sur le dépliement thermique de la protéine et de son domaine de liaison ainsi que sur son équilibre d'auto-association a été réalisé (Usobiaga et al., 1996) (Medrano et al., 1996). De ces expériences, nous pouvons tirer les observations suivantes : l'autolysine complète LytA existe à plus de 80% sous forme de dimère, même en absence de choline. Ces dimères sont fortement stabilisés par l'interaction avec la choline (Varea et al., 2000). La forme dimérique, probablement allongée, faciliterait la diffusion de l'enzyme dans la structure du peptidoglycan. Cette dimérisation pourrait aussi faciliter l'interaction simultanée de l'enzyme avec deux acides téichoïques et permettrait à l'enzyme de progresser en « marchant » le long du septum par un mécanisme de liaison/relarguage en utilisant les modules C-terminaux des deux sous-unités sans se détacher complètement du substrat.

La composition en structures secondaires de ClytA, déterminée par dichroïsme circulaire dans les UV lointains et spectrométrie infrarouge à transformée de Fourier, montre que ce domaine se compose de 17% d'hélices α, 51% de feuillets β, 20% de *turns* et de 12% de segments désordonnés (Medrano et al., 1996).

La liaison à la choline ne provoque pas de grands changements dans la composition en structures secondaires. Il n'y a donc pas de remaniements structuraux importants après liaison à la choline.
Les spectres de dichroïsme circulaire dans les UV proches suite à l'ajout de choline montrent des changements attribuables à des modifications dans l'environnement des tryptophanes et tyrosines. Ces résultats suggèrent que les interactions cation-π entre la choline et les chaînes latérales des résidus aromatiques hautement conservés dans le module C-terminal pourraient intervenir dans le processus de reconnaissance du ligand, comme cela a été décrit pour l'acétylcholineestérase (Harel et al., 1993).
La spectroscopie par dichroïsme circulaire en présence ou absence de choline met en évidence l'existence de deux types de sites de liaison à la choline présentant des affinités différentes. La saturation par la choline des sites de haute affinité est impliquée dans la dimérisation et l'augmentation subséquente d'affinité pour le substrat. Au contraire, la saturation des sites de faible affinité nécessite des concentrations importantes en choline (plus de 10 mM).
En résumé, ces expériences nous apprennent que le domaine de liaison à la choline ClytA est principalement composé de brins β et qu'il n'y a pas de remaniements structuraux importants suite à l'ajout de choline. Par contre, il semble y avoir une contribution des résidus aromatiques dans la liaison enzyme-ligand. De plus, il existerait deux types de sites de liaison à la choline, répartis sur les 6 *repeats*, et présentant des affinités différentes pour la choline.
Afin de mieux comprendre la structure du domaine de liaison ClytA et d'attribuer des fonctions potentielles différentes aux séquences répétées le composant, deux expériences de délétions progressives ont également été réalisées au départ de l'extrémité C-terminale de la protéine.

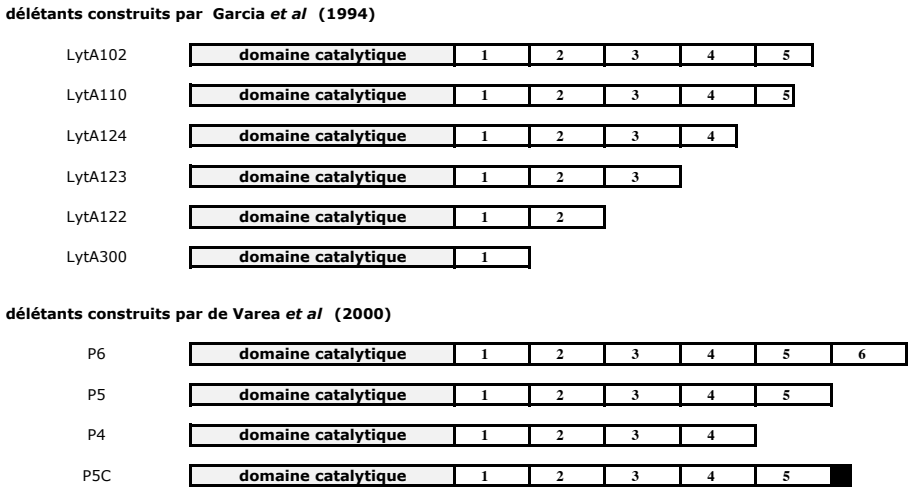


Figure 16 Représentation schématique des délétants créés pour analyser la structure du domaine de liaison ClytA et attribuer des rôles potentiels aux différents *repeats*. Les numéros correspondent aux différents *repeats* encore présents. La zone en noir représente la queue C-terminale de onze résidus.

Dans un premier temps, Garcia et ses collaborateurs ont construits six délétants (figure 16) (Garcia et al., 1994). Ces délétions concernent les *repeats* 4, 5 et 6 ainsi que la queue C-terminale. Afin de caractériser les enzymes tronquées, celles-ci ont été exprimées dans *Escherichia coli*. Puis, des extraits bruts de la bactérie ont été déposés sur des filtres de DEAE (analogue structural de la choline) et des élutions progressives dans un gradient de NaCl puis à la choline 2% ont été réalisées. Ce type d'expérience permet de mesurer l'affinité résiduelle des délétants pour le DEAE et donc, de façon indirecte, leur affinité pour la choline. La récupération de l'amidase tronquée est contrôlée par migration de chaque fraction éluée sur gel SDS-page et par mesure de l'activité enzymatique résiduelle. Les délétions progressives du domaine de liaison à partir de son extrémité C-terminale permettent de tirer les conclusions suivantes :

- **l'amidase doit conserver au moins quatre *repeats* pour reconnaître efficacement la choline.** La perte d'un *repeat* supplémentaire diminue drastiquement l'activité enzymatique et l'affinité de liaison
- les délétants pour la queue C-terminale et les *repeats* 5 et 6 sont plus sensibles à l'inhibition par la choline libre que LytA. Cette observation corrèle l'existence de plusieurs sites de liaison répartis sur tout le domaine. En effet, LytA possédant plus de sites de liaison nécessiterait des concentrations en choline plus importantes pour être totalement inhibée
- la comparaison de l'activité de LytA et des formes tronquées montre que la diminution d'activité se fait en trois étapes majeures. La première étape correspond à la perte de la queue C-terminale abolissant le processus de conversion. La seconde est liée à la perte du *repeat* 6 produisant une

diminution de l'activité enzymatique tout en maintenant la capacité de liaison à la choline. La dernière étape correspondant à la délétion du *repeat* 4 cause une perte de capacité de liaison à la choline. Ainsi, l'acquisition de *repeats* et de la queue C-terminale pourrait être considéré comme un avantage évolutif puisque l'addition de ces motifs augmente l'activité hydrolytique et permet une régulation.

Dans une deuxième expérience de délétions à partir de l'extrémité C-terminale, Varea et ses collaborateurs ont investigué l'importance des *repeats* 4, 5, 6 et de la queue C-terminale sur la capacité de liaison , la structure native et la stabilité de la protéine par des expériences de spectroscopie en dichroïsme circulaire, de spectrométrie infrarouge à transformée de Fourier et de calorimétrie différentielle à balayage (Varea et al., 2000). Ils en retirent les conclusions suivantes :

- le *repeat* 5 est impliqué dans la stabilisation du domaine de liaison et du domaine catalytique de l'amidase. Il affecte la stabilité des sites de liaison à la choline. Ce *repeat* semble participer à un *pathway* coopératif hypothétique, permettant la communication moléculaire entre les deux modules de l'amidase,
- le *repeat* 6 semble jouer un rôle structural important, agissant comme "*spacer*" physique permettant aux contacts natifs d'être formés par la queue. En effet, l'ajout de la queue C-terminale après le *repeat* 5 ne permet pas d'observer la formation de dimères. Ce *repeat* a donc une fonction plus topologique que de stabilisation dans l'organisation structurale, permettant probablement une orientation correcte de la queue C-terminale,
- la queue C-terminale joue un rôle clef dans la formation des dimères, la différenciation des sites de haute et de faible affinité et dans la stabilisation de la région C-terminale.

De tous ces résultats, nous pouvons retenir les éléments suivants pour la conception d'un *tag* d'affinité au DEAE :

- il existe des sites de liaison répartis sur l'ensemble de la séquence du domaine C-terminal puisque des protéines tronquées possédant de moins en moins de *repeats* présentent une affinité de plus en plus faible pour la choline,
- dans l'amidase native, la queue C-terminale permet de différencier des sites de liaison de haute affinité et de faible affinité. Cette particularité est cependant abolie par la délétion de cette extrémité C-terminale,
- l'amidase doit conserver au moins quatre *repeats* pour reconnaître efficacement la choline. La perte d'un *repeat* supplémentaire diminue drastiquement l'affinité de liaison,
- les résidus aromatiques que l'on retrouve de façon récurrente dans la séquence et qui apparaissent dans des consensus de séquence impliquant les *repeats* des CBP semblent impliqués dans la liaison à la choline, comme c'est le cas pour l'acétylcholinestérase.

3. Stratégies de conception d'un *tag* de purification

3.1. Trois approches envisageables pour créer un tag

La première méthode consiste à utiliser des banques peptidiques ou de domaines protéiques et à sélectionner, par étapes successives, un ou plusieurs candidats présentant une affinité pour un ligand particulier. Les banques utilisées sont généralement présentées en surface de bactéries ou de phages (technologies du *phage-display* ou du *bacterial display*) (Lu et al., 1995) (Smith, 1985). Ces techniques ont été appliquées avec succès, par exemple, pour la création du *tag* Strep (Schmidt and Skerra, 1993). Parmi les avantages de cette approche, nous pouvons citer la possibilité de tester un très grand nombre de candidats et le lien physique direct qui existe entre la séquence génétique codant pour le candidat et le candidat peptidique. Dans le cadre de la création d'un *tag* de purification à usage industriel, le désavantage majeur de l'utilisation de ces technologies est la nécessité d'obtenir des licences d'exploitation, ces technologies étant protégées par des brevets.
Une deuxième méthode consiste à réaliser diverses découpes d'un domaine complet présentant l'affinité recherchée, afin de sélectionner la séquence peptidique la plus courte possible possédant toujours une affinité pour le DEAE. Cette méthode, plus laborieuse, ne permet pas de tester un grand nombre de candidats par des méthodes simples et peu onéreuses. Elle peut cependant constituer un point de départ, à partir duquel le fragment peptidique sélectionné sera soumis à une mutagenèse rationnelle ou aléatoire, visant à améliorer certaines qualités (affinité, solubilité, stabilité) de ce fragment. La combinaison de ces deux premières approches a été appliquée pour les peptides dérivés du domaine Z de la protéine A (Graslund et al., 2002) (Gunneriusson et al., 1999).
Une troisième méthode peut être envisagée. Elle consiste à combiner des approches bioinformatiques visant à identifier des résidus-clés dans la séquence peptidique et/ou des motifs structuraux essentiels dans sa conformation tri-dimentionnelle, dans le but de définir le motif responsable de la propriété recherchée. Une fois cette étape réalisée, on peut alors envisager la synthèse d'une séquence originale reprenant les éléments importants mis en évidence.

3.2. Les outils bioinformatique utilisables pour la conception d'un tag de purification

L'accumulation de données produites dans le domaine de la biologie moléculaire et de données provenant du séquençage de génomes a poussé au développement de nouvelles techniques permettant une analyse rapide de ces informations. C'est ainsi qu'il y a quelques années est née une nouvelle

discipline alliant la biologie à l'informatique : la bioinformatique. La bioinformatique est la discipline de l'analyse de l'information biologique, en majorité sous la forme de séquences génétiques et de séquences et structures protéiques.
Les programmes d'analyse de l'information biologique actuels couvrent de vastes champs à la fois différents et complémentaires.

Au niveau de l'analyse protéique, nous pouvons distinguer deux types de prédiction qui sont à la fois liés et complémentaires : la prédiction de conformation et la prédiction de fonction.

La prédiction de la fonction d'une protéine inclut notamment la recherche, dans des banques de données de séquences homologues de fonction connue. Cette recherche peut se faire par des programmes d'alignements de séquences pairés ou multiples. Un autre outil intéressant consiste à rechercher des motifs ou des séquences répétées caractéristiques d'une fonction particulière. Une prédiction de signaux tels que des sites protéolytiques liés à la présence de peptides signaux peuvent aider à prédire la localisation de la protéine dans la cellule. Des sites de glycosilation ou phosphorylation,… constituent d'autres outils de prédiction. Au niveau de la séquence de la protéine d'intérêt, la recherche de régions hydrophobes ou hydrophiles, de groupes de résidus chargés, de périodicité dans la distribution des différents résidus sont des indications utiles. Enfin, la prédiction de structures secondaires, la topologie et localisation d'hélices transmembranaires, de régions *coiled coil*, ;…peuvent intervenir dans la recherche d'une fonction.
La prédiction de structures secondaires constitue également une des premières démarches dans la prédiction de la structure d'une protéine. L'attribution d'une conformation particulière se fait aussi par recherche dans les banques de données de sous-structures ou de domaines connus.
Ces deux premiers types d'analyses rentrent dans le vaste domaine de la modélisation de la structure des protéines. Ce dernier peut schématiquement se diviser en trois types d'approche :
- la modélisation par homologie. Cette approche est basée sur l'observation que la nature met en jeu un nombre relativement restreint de motifs structuraux pour élaborer l'immense diversité du protéome (Lambert et al., 2003). Elle s'effectue en quatre étapes principales. La première est la recherche d'une ou plusieurs séquences de structure connue (appelée(s) *template(s)*), similaire(s) à la séquence d'intérêt. Pour une bonne qualité de modélisation, le pourcentage d'identité entre le(s) *template(s)* et la séquence d'intérêt doit être supérieure à 35%. Dans la deuxième étape, la séquence d'intérêt est alignée aux séquences des *templates*. Cette étape est cruciale car elle est la principale source d'erreur de cette technique de modélisation. Ensuite, les coordonnées tri-dimensionnelles des résidus du (des) *template(s)* sont attribués aux résidus de la séquence d'intérêt les portions de la séquence

d'intérêt non alignées doivent être modélisées. Enfin, la structure prédite pour la séquence d'intérêt doit être vérifiée d'un point de vue énergétique et géométrique. La modélisation par homologie est une des premières techniques de modélisation mise au point et reste, à ce jour, une des plus précises.
- la reconnaissance d'un *fold* protéique (*fold recognition*). On utilise généralement cette approche lorsqu'on ne trouve pas de séquences de structure connue, homologues à la séquence d'intérêt. Dans ce cas, le concept de base sous-jacent est que des protéines présentant des séquences et des fonctions très différentes peuvent quand même adopter des *folds* très similaires (Lemer et al., 1995). On distingue deux types de méthode de reconnaissance de *fold*. Le *threading* fait intervenir l'alignement de la séquence protéique d'intérêt aux séquences d'une banque de *folds* et l'évaluation de la compatibilité de la séquence avec la structure par une fonction d'énergie (Flockner et al., 1995). Le pseudo-*threading* décrit chaque *fold* de la banque comme une succession (ou un profil) de propriétés associées à chaque résidu de la structure. Cette méthode va donc tenter d'aligner les profils à la séquence à modéliser. Le meilleur alignement sert alors à prédire la structure tri-dimensionnelle que doit adopter la séquence d'intérêt.
- la modélisation *ab initio*. Dans ce cas, le concept de base est qu'une protéine adopte sa conformation native à son minimum d'énergie libre de Gibbs. On peut estimer cette énergie grâce à des fonctions empiriques (champs de force) décrivant l'ensemble des interactions subies par chaque atome d'une protéine. Actuellement, cette technique en est encore à ses balbutiements car les règles régissant le repliement des protéines ne sont pas encore toutes élucidées et leur simulation requiert des temps de calcul encore inaccessibles.

La bioinformatique couvre encore de nombreux autres champs tels que :

- l'analyse des séquences nucléotidiques. Cette dernière peut permettre la détermination des régions codantes et l'attribution systématique de la structure de toutes les protéines d'un génome (génomique structurale) ou l'analyse de régions non codantes (recherche de promoteurs, de régions régulatrices, des signaux de début et fin de transcription, des signaux de début et fin de traduction, de régions de contrôles communes à plusieurs gènes, …),
- l'organisation, dans des bases de données relationnelles de quantités impressionnantes d'informations provenant des programmes de séquençage de génomes. Cette masse de données pose de sérieux défis pour assurer la validation des informations, leur mise à jour, leurs recoupements, leur stockage, leur accessibilité, leur protection, …
- la modélisation des voies métaboliques,…

3.3. L'approche bioinformatique dans le cadre de la création d'un tag de purification par affinité pour le DEAE-Sépharose

Comme nous l'avons déjà mentionné, si l'utilisation de banques aléatoires ou semi-aléatoires avec présentation en surface de phages ou de bactéries constitue une voie intéressante pour la conception d'un *tag* de purification par affinité, le concepteur se heurte cependant au problème d'obtention de licences d'exploitation liées à l'utilisation de ces technologies.
Notre projet est réalisé en partenariat avec la firme pharmaceutique GlaxoSmithKline Biologicals, nous devons tenir compte de ces impératifs dans le choix de notre stratégie.
Afin d'éviter tout problème d'exploitation ultérieure du *tag*, nous avons décidé de suivre une approche bioinformatique. L'emploi de certains outils de bioinformatique nous semble réaliste pour la création d'un *tag* de purification au départ d'un domaine de liaison à la choline.
En effet, d'une part, des expériences montrent que les séquences répétées responsables de la liaison à un ligand peuvent être déterminées, notamment par des alignements de leur séquence (Wren, 1991) (Giffard and Jacques, 1994). On espère ainsi mettre en évidence des résidus bien conservés jouant potentiellement un rôle essentiel dans la propriété de liaison recherchée.
D'autre part, des expériences de délétions progressives sur le domaine C-terminal des protéines LytA et PspA de *Streptococcus pneumoniae* montrent qu'il n'est pas indispensable de garder le domaine de liaison complet pour conserver une affinité pour la choline (Garcia et al., 1994). On peut donc raisonnablement penser qu'il est possible de diminuer la taille des domaines de liaison naturels en diminuant le nombre de motifs répétés et en améliorant ces motifs répétés sur base de caractéristiques dégagées, par exemple, par des expériences d'alignements.
Au début de ce travail, aucune structure de domaine de liaison à la choline n'était connue. Nous avons donc fait appel aux techniques de modélisation par homologie ou de reconnaissance de *fold* pour essayer d'établir un modèle, même approximatif de ces domaines. Ce modèle pouvait ensuite permettre de définir le nombre minimum de motifs répétés nécessaires pour la liaison à la choline et le nombre de molécules de choline qu'un domaine complet peut lier.

OBJECTIF ET STRATEGIES

Comme nous l'avons commenté dans l'introduction, la mise au point de systèmes de purification de protéines répond à une demande croissante et importante du monde industriel. Bien qu'il existe déjà de nombreuses techniques de purification, l'industrie est toujours à la recherche de nouveaux procédés soit répondant à des critères bien particuliers tels que le type de gel chromatographique utilisé, soit les affranchissant des licences d'exploitations des systèmes brevetés existants.
Dans ce contexte, le but de notre travail est la création d'un nouveau *tag* de purification par affinité pour le gel DEAE-Sépharose.
Ce gel est une matrice échangeuse d'anion faible, couramment utilisée au cours de chromatographies par échange d'ions. Le DEAE-Sépharose permet à toute protéine chargée négativement de s'adsorber sur la colonne, l'élution des différents polypeptides se réalisant ensuite par application d'un gradient de sels.
La conception de ce *tag* est rendue possible par le fait qu'il existe dans la nature des domaines de liaison à la choline. La choline étant un analogue structural du DEAE, nous espérons qu'un *tag* dérivé d'un domaine de liaison à la choline présentera une affinité spécifique pour le DEAE-Sépharose, l'élution de la protéine d'intérêt se réalisant par compétition entre le *tag* et une solution de choline, indépendamment de la charge de la protéine à purifier.
Les domaines de liaison à la choline sont constitués de séquences répétées d'une vingtaine d'acides aminés appelés *repeats*. La séquence des *repeats* n'est pas conservée à certaines positions et le nombre de *repeats* constituant un domaine de liaison est également variable (de quatre à seize *repeats*).
Pour créer un nouveau *tag* de purification, nous avons décidé d'utiliser comme point de départ la séquence du domaine de liaison à la choline de l'amidase LytA de *Streptococcus pneumoniae*, l'une des plus étudiées dans la littérature.
Son domaine se compose de six séquences répétées. Des expériences de délétions progressives sur le domaine LytA montre que quatre *repeats* suffisent pour conserver une affinité pour le DEAE. On peut donc raisonnablement penser qu'il est possible de créer un *tag* assez court au départ de ce domaine. De plus, des expériences d'alignements sur les *repeats* de domaines de liaison à la choline ou à des sucres permettent de dégager un consensus de *repeats* mettant en évidence des résidus jouant un rôle important dans cette propriété de liaison.
Les données de la littérature que nous venons de passer en revue nous ont permis d'élaborer deux stratégies parallèles, l'une bioinformatique, l'autre expérimentale.
Pour caractériser la séquence des *repeats* des domaines de liaison à la choline, nous avons recherché toutes les séquences de domaines de liaison à la choline connus et aligné leurs *repeats* à l'aide de plusieurs algorithmes, permettant ainsi d'élaborer un consensus de séquence. Aucune structure de ce type de domaine n'étant connue au début de ce projet, nous avons tenté de

construire un modèle tri-dimensionnel par les techniques de modélisation par homologie ou de reconnaissance de *fold*s. Nous espérions que les résultats obtenus par ces deux types d'analyse nous permettraient de mieux définir les résidus importants pour la liaison à la choline et de les situer dans l'espace. Sur base des données escomptées, l'étape suivante consistait à définir des séquences peptidiques semi-aléatoires reprenant les informations structurales et de séquence nécessaires pour obtenir une affinité pour le DEAE permettant la purification ultérieure de protéines d'intérêt sur cette matrice.
En parallèle, nous avons développé une approche expérimentale complétant les premières données de délétions progressives réalisées par Garcia (Garcia et al., 1994). A cette fin, toujours au départ du domaine de liaison C-LytA, nous avons créé des versions tronquées supplémentaires du domaine de liaison et testé leur capacité de liaison sur gel DEAE-Sépharose.

RESULTATS

PARTIE I : ANALYSES BIOINFORMATIQUES DES DOMAINES DE LIAISON A LA CHOLINE

I.1 Analyse des caractéristiques physico-chimiques des séquences des domaines de liaison à la choline

En l'absence de toute information sur la structure, une façon de caractériser les motifs répétés responsables de la liaison à la choline est de rechercher dans les domaines complets la répartition particulière de certains acides aminés. En effet, des groupes d'acides aminés hydrophobes ou chargés, placés de façon périodique tout au long du domaine peuvent contribuer à définir les éléments du *repeat* indispensables pour la liaison à la choline.
Avant de débuter ce travail d'analyse de séquences, nous avons d'abord recherché dans les banques de données toutes les protéines présentant un domaine similaire à celui de l'amidase LytA. Pour ce faire, nous avons soumis la séquence du domaine de liaison C-LytA au programme de recherche PSI-BLAST (Altschul et al., 1997), en utilisant les paramètres par défaut et la banque de données non redondante (NR, version 2000). Sur base des résultats de cette recherche, 19 domaines de liaison ont été sélectionnés (tableau 4). Chaque séquence de domaine a ensuite été soumise au programme d'analyses statistiques de séquences protéiques SAPS (Brendel et al., 1992). Aucune particularité dans la distribution des résidus ou des structures secondaires n'a pu être relevée sur la séquence du domaine de liaison CLytA.
Nous en concluons que la liaison à la choline ne repose pas simplement sur la présence de groupes d'acides aminés hydrophobes ou chargés mais que d'autres caractéristiques physico-chimiques ou structurales interviennent.

Protéine	Fonction	Nombre de motifs répétés dans le domaine de liaison à la choline	Espèce bactérienne ou espèce phagienne
CbpA (Rosenow et al, 1997)	adhésine	10 motifs	*Streptococcus pneumoniae*
CPL1(Garcia et al, 1987)	hydrolase de la muréine	6 motifs	phage Cp-1
CPL9 (Garcia et al, 1988)	hydrolase de la muréine	6 motifs	phage Cp-9
CspA (Sanchez-Beato et al, 1995)	hydrolase de la muréine	4 motifs	*Clostridium beijerinckii*
CspB (Sanchez-Beato et al, 1996)	fonction inconnue	6 motifs	*Clostridium beijerinckii*
CspC (Sanchez-Beato et al, 1996)	fonction inconnue	4 motifs	*Clostridium beijerinckii*
CspD (Sanchez-Beato et al, 1996)	fonction inconnue	4 motifs	*Clostridium beijerinckii*
EJL (Diaz et al, 1992)	hydrolase de la muréine	6 motifs	phage EJ-1
HBL3 (Romero et al, 1990)	hydrolase de la muréine	6 motifs	phage HB-3
LytA (Ronda et al, 1987)	hydrolase de la muréine	6 motifs	*Streptococcus pneumoniae*
LytB (Garcia et al, 1999)	hydrolase de la muréine	15 motifs	*Streptococcus pneumoniae*
LytC (Garcia et al, 1999)	hydrolase de la muréine	11 motifs	*Streptococcus pneumoniae*
PbcA (Cheng et al, 2000)	C3-binding protein	4 motifs	*Streptococcus pneumoniae*
PcpA (Sanchez-Beato et al, 1998)	fonction inconnue	13 motifs	*Streptococcus pneumoniae*
PcpC (non publié, numéro d'accès CAB04760 dans Swissprot)	fonction inconnue	16 motifs	*Streptococcus pneumoniae*
PAL (non publié, numéro d'accès CAB07986 dans Swissprot)	hydrolase de la muréine	3 motifs	phage Dp-1
PspA (Yother et al, 1992)	facteur de virulence	10 motifs	*Streptococcus pneumoniae*
PspC (Brooks-Walter et al, 1999)	fonction inconnue	11 motifs	*Streptococcus pneumoniae*
SpsA (Hammerschmidt et al, 1997)	fonction inconnue	4 motifs	*Streptococcus pneumoniae*

Tableau 4 Liste des protéines possédant un domaine de liaison à la choline. Ces protéines ont été sélectionnées dans les banques de données par le programme PSI-BLAST après soumission de la séquence du domaine de liaison ClytA.

I.2. Etablissement d'un consensus des séquences répétées composant les domaines de liaison à la choline

I.2.1. Définition des motifs répétés

Tous les domaines de liaison n'ayant pas été caractérisés, la première partie de ce travail a consisté à redéfinir les *repeats* contenus dans chaque domaine. L'observation des motifs déjà décrits dans la littérature permet de tirer les caractéristiques suivantes :

- grande conservation d'un triplet d'acides aminés aromatiques
- la distance entre deux triplets d'aromatiques est de l'ordre de 20 acides aminés.

De façon arbitraire, nous avons donc découpé les domaines de liaison à la choline en considérant qu'un motif répété consistait en un triplet d'aromatiques entouré de 9 acides aminés en amont et en aval.
Dans la majorité des séquences analysées, cette définition ne pose pas de problème puisque les *repeats* redéfinis se superposent approximativement aux *repeats* déjà décrits (Brooks-Walter et al., 1999) (Yother and White, 1994). Par contre, nos résultats ne concordent pas pour les protéines LytB et

LytC. D'après la littérature, ces protéines possèdent respectivement 15 *repeats* et 11 *repeats* (Garcia et al., 1999a) (Garcia et al., 1999b). En appliquant notre définition, nous trouvons respectivement 13 *repeats* et 6 *repeats*. Une analyse plus fine de la séquence permet de définir 5 motifs dégénérés supplémentaires pour LytC. Nous entendons par motif dégénéré un motif ne possédant pas un triplet d'aromatiques mais une paire d'aromatiques ou un motif présentant une variabilité dans le nombre d'acides aminés entourant le triplet.
De même, dans la littérature, les protéines PcpA et PcpC possèdent respectivement 13 motifs dont un dégénéré et 16 motifs dont 12 dégénérés (Sanchez-Beato et al., 1998). En appliquant notre définition, nous retenons uniquement 12 motifs pour PcpA et 5 motifs dont 2 dégénérés pour PcpC.
Enfin, pour les domaines de liaison non caractérisés des protéines PAL et PbcA, nous trouvons 4 motifs répétés.
En reprenant l'ensemble des domaines de liaison, 133 *repeats* ont ainsi été définis. Nous reviendrons plus tard sur cette définition arbitraire du *repeat*.

I.2.2. Alignements multiples sur les 133 motifs répétés

Les *repeats* ont été soumis aux programmes d'alignements multiples Match-Box (Depiereux and Feytmans, 1992) et ClustalW (Thompson et al., 1994).
Une première analyse des résultats met en évidence des motifs répétés dont la séquence, généralement plus courte que celle des autres *repeats*, diminue la qualité de l'alignement en diminuant la taille de la boîte (Match-Box) ou en insérant de grands gaps (ClustalW).
Ces motifs répétés correspondent aux cinq *repeats* dégénérés définis pour LytC et aux deux motifs dégénérés de PcpC.
Afin d'optimiser les alignements, ces sept *repeats* ont été écartés, ramenant le nombre de motifs répétés à 126.

I.2.3. Alignements multiples sur les 126 repeats

Les 126 motifs répétés restant ont été soumis aux programmes d'alignements multiples Match-Box (Depiereux and Feytmans, 1992), ClustalW (Thompson et al., 1994), Block-Maker (utilisation des programmes MOTIF et GIBBS) (Henikoff et al., 1995) et Dialign2 (Morgenstern, 1999). Nous obtenons donc au total cinq alignements.
Pour chaque alignement, la fréquence relative de chaque acide aminé à chaque position a été calculée. Puis les acides aminés ont été regroupés en 8 classes et la fréquence relative de chaque classe à chaque position de l'alignement a été calculée.
Les huit classes ont été définies sur base des propriétés physico-chimiques et/ou structurales des acides aminés (tableau 5).

	Propriété physico-chimique ou structurales des résidus de la classe	**résidus présents dans la classe**
classe n° 1	hydrophobes	alanine (A), leucine (L) isoleucine (I), méthionine (M) valine (V)
classe n°2	aromatiques	tryptophane (W), tyrosine (Y) phénylalanine (F)
classe n°3	négatifs ou acides	acide aspartique (D) acide glutamique (E)
classe n°4	positifs ou basiques	lysine (K), arginine (R) histidine (H)
classe n°5	neutres polaires	sérine (S), thréonine (T), asparagine (N) et glutamine (Q)
classe n°6	atome d'hydrogène comme chaine latérale flexibilité conformationnelle importante	glycine (G)
classe n°7	formation éventuelle de ponts disulfures	cystéine (C)
classe n°8	AA cyclique imposant des contraintes conformationnelles	proline (P)

Tableau 5 Définition de huit classes d'acides aminés en fonction des propriétés physico-chimiques et / ou structurales de leur chaîne latérale.

L'importance relative de chaque classe d'acides aminés aux diverses positions de l'alignement a été définie arbitrairement comme suit (tableau 33 en Annexe 1) :

- lorsqu'au moins 60 % des acides aminés présents à une position appartiennent à la même classe, seule cette classe est retenue dans le consensus,
- lorsque deux classes différentes, présentant des caractéristiques physico-chimiques compatibles, totalisent au moins 70 % des acides aminés à une position et que cette caractéristique se retrouve pour les cinq alignements obtenus, les deux classes apparaissent à cette position dans la définition du profil. Nous entendons par caractéristiques physico-chimiques compatibles les groupes polaires-chargés ou aromatiques/hydrophobes,
- lorsque deux classes d'acides aminés totalisent 70% de résidus pour certains programmes d'alignement mais pas pour d'autres (discordance entre les cinq alignements obtenus), un X est noté à cette position dans le consensus (position variable),
- lorsque plusieurs classes présentant des caractéristiques physico-chimiques différentes apparaissent à une position dans des proportions similaires, un X est également noté à cette position dans le consensus.

En final, on obtient donc pour chaque programme d'alignement et à chaque position la ou les classes d'acides aminés les plus fréquentes. Ces résultats obtenus par les cinq programmes sont concordants puisqu'ils mettent chaque

fois en évidence la ou les mêmes classes de résidus à chaque position (tableau 33 annexe 1).
Ensuite, en reprenant les cinq consensus obtenus, un consensus général a été dégagé (tableau 6).

position	1	2	3	4	5	6	7	8	9	10
classe de résidus	pol	G	arom	phobe	ch/pol	X	ch/pol	G	X	arom
résidu le plus fréquemment retrouvé dans une classe	T		W	L/V						W
position	11	12	13	14	15	16	17	18	19	20
classe de résidus	arom	arom	phobe	ch/pol	X	ch/pol	G	X	phobe	phobe
résidu le plus fréquemment retrouvé dans une classe	Y	Y	L						M	A

Tableau 6 Consensus obtenu après alignement des 126 motifs répétés par les programmes d'alignement multiple ClustalW, Dialign2, Blockmaker et Match-Box.
Pol : classe des résidus polaires
G : glycine
Arom : classe des résidus aromatiques
Phobe : classe des résidus hydrophobes
Ch/pol : classes des résidus chargés ou polaires
X : position variable
Les lettres majuscules correspondent aux symboles des 20 acides aminés. Une lettre inscrite à la place d'une classe d'acides aminés signifie que cet acide aminé est majoritaire à cette position du consensus. Une lettre inscrite sous une classe d'acides aminés signifie qu'il s'agit de l'acide aminé le plus fréquemment observé à cette position du consensus pour cette classe d'acides aminés.

Afin de mieux caractériser les positions X obtenues pour les consensus des divers programmes d'alignements multiples, deux approches ont été suivies.
La première consiste à calculer la probabilité d'observer une classe d'acides aminés à une position particulière, compte tenu de la fréquence d'utilisation des acides aminés (fréquence attendue) dans les protéines de liaison à la choline de *Streptococcus pneumoniae* et *Clostridium beijerinckii*. Cette fréquence attendue a été calculée en reprenant les domaines d'activité catalytique de ces protéines et en les soumettant au programme d'analyse statistiques des séquences protéiques SAPS. Le pourcentage de chaque acide aminé a ensuite été divisé par le nombre total de résidus afin d'obtenir la fréquence relative de cet acide aminé dans les protéines possédant un domaine de liaison à la choline. Ensuite, les fréquences relatives des acides aminés ont été additionnées par classe, permettant d'obtenir la fréquence relative d'utilisation des acides aminés présentant les mêmes caractéristiques physico-chimiques (tableau 34 annexe 1).
En comparant la fréquence attendue d'une classe d'acides aminés au nombre de fois que l'on observe réellement cette classe à une position de

l'alignement, on calcule par une loi binomiale une probabilité d'observation. Plus cette probabilité est petite, plus la présence (ou absence) d'une classe d'acides aminés à une position du consensus est significative. En effet, la probabilité est très petite aussi bien pour les acides aminés sous-représentés que pour les acides aminés sur-représentés. Un exemple de ce type d'analyse est donné au tableau 35 de l'annexe 1.
En appliquant cette analyse de probabilité à toutes les positions variables X, un consensus recorrigé a été établi (tableau 7).

position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
consensus 2 après calcul de probabilité	pol T	G	arom W	phobe L/V	ch/pol	**acide**	ch/pol	G	**polaire**	W

position	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
consensus 2 après calcul de probabilité	Y	Y	phobe L	ch/pol	**polaire**	ch/pol	G	**polaires**	phobe M	phobe A

Tableau 7 Consensus établi après calcul de probabilité aux positions variables X établies par le premier consensus. Les symboles et abréviations utilisés sont identiques à ceux utilisés et expliqués au tableau 6.

Une deuxième approche consiste à classer les acides aminés présents à une position variable X en fonction du volume de leur chaîne latérale. En effet, d'un point de vue structural, la présence de petits acides aminés à des positions précises peut permettre l'adoption d'une conformation particulière, indépendamment des caractéristiques physico-chimiques de ces acides aminés. Pour ce faire, nous avons utilisé comme volumes de référence les volumes des chaînes latérales définis par Krigbaum (Krigbaum and Komoriya, 1979). Nous avons répartis arbitrairement les 20 acides aminés en quatre groupes décrits au tableau 8.

	volume de la chaîne latérale ($Å^3$)	résidus concernés
groupe 1	0 - 40	A, D, S, G
groupe 2	41 - 70	N, C, P, E, T, H
groupe 3	71 - 100	L, K, M, Q, I, V
groupe 4	> 100	R, F, Y, W

Tableau 8 Classification des acides aminés en fonction du volume de leur chaîne latérale. Les lettres correspondent aux symboles des 20 acides aminés.

A nouveau, nous avons calculé le nombre d'acides aminés de chaque groupe à chaque position variable et ce, pour chaque programme d'alignements multiples. Le tableau 36 en Annexe 1 reprend les résultats obtenus. Nous pouvons observer que les deux dernières positions variables présentent un grand pourcentage de petits acides aminés. Il sera donc intéressant de tenir compte de cette caractéristique dans le consensus final.

I.2.4. Classification des motifs répétés

Tous les motifs répétés composant un domaine de liaison à la choline ne semblent pas nécessaires pour établir cette liaison. En effet, des expériences de délétions progressives sur la protéine LytA montrent que sur les 6 *repeats* du domaine, 4 sont suffisants pour avoir une affinité détectable pour la choline (Garcia et al., 1994). Les motifs répétés n'intervenant pas ou peu dans la fonction de liaison pourraient théoriquement subir moins de pression de sélection et donc présenter plus de variabilité au niveau de leur séquence. De tels *repeats* pris en compte dans les alignements pourraient affaiblir le consensus.
Il nous a donc paru intéressant d'essayer de classer les *repeats* de différentes façons afin de voir si certains groupes de motifs ne présentaient pas une variabilité très importante, nous permettant alors de les exclure des alignements. De plus, ces classifications suivies d'alignements multiples aboutissent également à la définition de consensus. La comparaison de ces consensus au consensus déjà obtenu par l'alignement des 126 *repeats* permettra peut-être de confirmer ou d'infirmer les positions variables X déjà définies ainsi que les positions pour lesquelles une classe d'acides aminés semble être prépondérante.

I.2.4.1. Classification des repeat*s par degré de similarité*

Une première façon de classer les motifs répétés est de les soumettre au programme d'alignements multiples ClustalW (Thompson et al., 1994). Ce dernier crée un "arbre guide", pouvant être grossièrement assimilé à une classification phylogénique des *repeat*s. En fonction de cet arbre, 11 modules contenant un nombre variable de motifs ont été créés (tableau 37 Annexe 1).
Chaque module a ensuite été soumis au programme d'alignement multiple Match-Box puis, les résultats d'alignements ont été traités selon la méthode décrite précédemment afin d'établir un consensus (Depiereux and Feytmans, 1992). Des calculs de probabilités effectués sur les positions variables X de ce consensus permettent d'essayer de préciser la ou les classes d'acides aminés importantes pour la fonction de liaison à la choline à certaines de ces positions.

Le consensus final obtenu pour cette classification phylogénique est repris au tableau 9.

Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Consensus 1	**pol X**	**G X**	arom W	**X**	ch/pol	**X**	ch/pol	**X G**	**X**	W
Consensus 2 après calcul de probabilité	**X**	**X**	arom W	**X**	ch/pol	**X**	ch/pol	**G**	**X**	W

Position	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Consensus 1	Y	Y	phobe	ch/pol	**X**	ch/pol	G	**X**	**phobe (x)**	**phobe (x)**
Consensus 2 après calcul de probabilité	Y	Y	phobe	ch/pol	**X**	ch/pol	G	**phobe arom**	**phobe**	**phobe**

Tableau 9 Consensus établi pour les 11 modules reflétant la répartition phylogénique des 126 *repeats*. Chaque module a été soumis au programme d'alignements multiples Match-Box afin d'obtenir un consensus par module (Depiereux and Feytmans, 1992). Ces 11 consensus ont ensuite été comparés de façon à dégager un premier consensus général (consensus 1). Enfin, des calculs de probabilité ont été réalisés aux positions variables pour mieux définir les résidus les plus représentés à ces positions. L'ensemble de ces résultats permet de définir un consensus général (consensus 2). Les symboles et abréviations utilisés sont identiques à ceux utilisés et expliqués au tableau 6.

I.2.4.2. Classification des motifs en fonction de leur position dans le domaine de liaison

Une autre idée de classification est de regrouper les motifs répétés en fonction de leur position dans le domaine de liaison. Les domaines de liaison à la choline possédant des longueurs très différentes (3 à 15 *repeats*), certains groupes seront peu représentés (par exemple les *repeats* 9, 10 et 11) alors que d'autres sont très bien représentés (par exemple, les *repeats* 1, 2 ou 3). Lors de l'établissement du consensus par cette méthode, nous aurons donc plus tendance à privilégier les résultats obtenus dans les groupes bien représentés et à moins tenir compte des résultats obtenus pour des groupes ne possédant, par exemple, que trois séquences de *repeat*.
Comme pour les analyses précédentes, les onze groupes formés sont soumis au programme d'alignement multiple Match-Box puis un consensus corrigé

par des calculs de probabilité est proposé (tableau 10) (Depiereux and Feytmans, 1992).

Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Consensus 1	**ch/pol X**	**G X**	arom	**X**	ch/pol	**X**	ch/pol	**X G**	**X**	W
Consensus 2 après calcul de probabilité	**pol**	**G**	arom	**phobe sauf repeats 1**	ch/pol	**X**	ch/pol	**G**	**X**	W

Position	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Consensus 1	Y	Y	phobe	ch/pol	**X**	ch/pol	G	**X**	**phobe (repeats 1)**	**phobe (X)**
Consensus 2 après calcul de probabilité	Y	Y	phobe	ch/pol	**X**	ch/pol	G	**X**	**phobe**	**phobe**

Tableau 10 Consensus obtenu après alignement des *repeats* classés en fonction de leur position dans le domaine de liaison. Onze groupes de *repeats* ont été définis par cette méthode. Ils ont ensuite été soumis au programme d'alignement multiples Match-Box et un consensus a été proposé pour chaque alignement (Depiereux and Feytmans, 1992). Ensuite, la comparaison des onze consensus a permis de dégager un premier consensus général (consensus 1) qui, après analyse de probabilité aux positions variables X, a permis de générer un deuxième consensus général (consensus 2).

I.2.4.3. Comparaison des consensus établis et établissement d'un consensus final

La comparaison des trois consensus généraux obtenus et l'analyse du volume de la chaîne latérale des acides aminés aux positions variables permettent l'établissement d'un consensus final (tableau 11).

	1	2	3	4	5	6	7	8	9	10
A	pol T	G	arom W	phobe L/V	ch/pol	acide	ch/pol	G	polaire	W
B	X	X	arom W	X	ch/pol	ch/pol	ch/pol	g	X	W
C	pol	g	arom	phobe sauf repeats 1	ch/pol	X	ch/pol	G	X	W
volume des chaînes latérales des AA						X			X	
consensus final	**pol/X**	**G/X**	**arom**	**phobe/X**	**ch/pol**	**X**	**ch/pol**	**G**	**X**	**W**

	11	12	13	14	15	16	17	18	19	20
A	Y	Y	phobe L	ch/pol	polaire	ch/pol	G	polaire	phobe M	phobe A
B	Y	Y	phobe	ch/pol	X	ch/pol	G	phobe arom	phobe	phobe
C	Y	Y	phobe	ch/pol	X	ch/pol	G	X	phobe	phobe
volume des chaînes latérales des AA					petit			petit		
consensus final	**Y**	**Y**	**phobe**	**ch/pol**	**petit AA**	**ch/pol**	**G**	**petit AA**	**phobe**	**phobe**

Tableau 11 Comparaison des consensus obtenus soit par alignements multiples des 126 *repeats* à l'aide de cinq programmes d'alignement (A), soit par classification des *repeats* en fonction de leur phylogénie puis alignement (B), soit par classification des *repeats* en fonction de leur position dans le domaine de liaison puis alignement (C). La taille des acides aminés (volume de leur chaîne latérale) aux positions 6, 9, 15 et 18 a été prise en considération pour la définition du consensus final.

Nous pouvons constater que certaines positions sont invariables quelle que soit la méthode de classification utilisée. Il s'agit notamment :

- de la position 3 : acide aminé aromatique, le plus souvent un tryptophane,
- des positions 5 et 7 : acide aminé polaire ou chargé,
- des positions 10, 11 et 12 : triplet d'acides aminés aromatiques, le plus souvent tryptophane, tyrosine, tyrosine,
- de la position 13 : acide aminé hydrophobe,
- des positions 14 et 16 : acide aminé chargé ou polaire,
- de la position 17 : une glycine,
- des positions 19 et 20 : acide aminé hydrophobe.

Les autres positions présentent des caractéristiques variables en fonction de la méthode de classification utilisée. Le tableau 11 montre que le consensus obtenu par alignement des modules phylogéniques s'écarte des consensus obtenus par les deux autres méthodes. Par exemple, on observe pour la position 1:

- une majorité de résidus polaires pour l'alignement des 126 *repeats*,

- une présence significative des résidus polaires après analyse de probabilité pour la classification par position,
- des résidus présentant des propriétés physico-chimiques variables pour la classification par modules phylogéniques, même après calcul de probabilités.

Même si les résultats obtenus pour ces positions montrent une certaine variabilité, ils ne seront pas totalement écartés lors de la conception de nos candidats, deux méthodes sur trois donnant des résultats convergents. De même, la taille des acides aminés aux positions 15 et 18 sera un facteur dont nous pourrons tenir compte dans la suite de notre travail.

> En résumé, nous avons établi un consensus de séquence des motifs répétés constituant les domaines de liaison à la choline. Ce consensus dérive de la comparaison de trois consensus établis sur base de trois démarches distinctes.
> Nous observons une nette conservation de résidus aromatiques et de glycines à certaines positions. Aux autres positions, il s'agit plutôt de classes d'acides aminés présentant des propriétés physico-chimiques similaires. Bien que ce consensus ne soit pas strict et ne permette donc pas d'établir une séquence *tag* candidate, il peut servir de base à l'élaboration de plusieurs candidats, en laissant certaines positions variables.

I.3. Prédiction des structures secondaires et recherche de structures tridimentionnelles connues, proches des domaines de liaison à la choline

I.3.1. Prédiction de structures secondaires sur l'entièreté des domaines de liaison à la choline et sur le consensus final

Comme nous l'avons signalé au paragraphe 1.2.1, nous avons arbitrairement défini les motifs répétés en considérant qu'un *repeat* consistait en un triplet de résidus aromatiques entouré, en amont et en aval, de 8 ou 9 autres résidus. Si cette définition arbitraire ne pose pas de problème pour rechercher par alignements multiples des résidus particulièrement conservés, elle ne tient aucunement compte de la présence éventuelle de structures secondaires.
Pour pallier ce problème, nous avons, dans un premier temps, soumis les 19 domaines de liaison au programme de prédiction de structures secondaires PSI-PRED, encore considéré actuellement comme un des programmes les plus performants dans ce domaine (Jones, 1999).
Les résultats montrent clairement que les domaines de liaison à la choline consistent en la succession de brins β et de coudes.
Nous avons ensuite soumis les motifs répétés pris individuellement au même programme afin de définir une position moyenne des brins β dans un *repeat*. Le tableau 12 reprend les résultats obtenus.

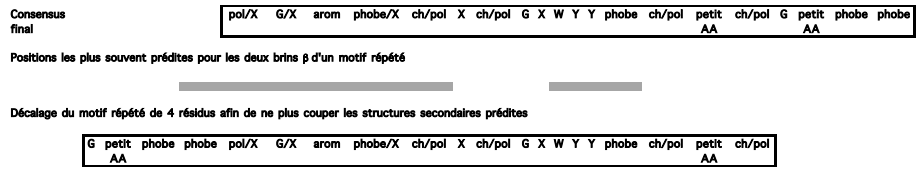


Tableau 12 Nouvelle définition d'un motif répété après analyse de la position des structures secondaires prédites. Les abréviations et symboles restent ceux définis précédemment.

Nous constatons qu'il y a en moyenne deux brins β prédits par motif. Ces résultats ont été confirmés par le programme de prédiction de structures secondaires PHD (Rost et al., 1994). La position des deux brins est telle qu'il nous faut décaler les acides aminés de quatre positions par rapport à la définition initiale du *repeat* pour ne pas couper le premier brin.

I.3.2. Recherche de structures tridimentionnelles connues, proches des domaines de liaison à la choline

Les alignements effectués aux paragraphes précédents permettent de mettre en évidence des acides aminés conservés jouant potentiellement un rôle important dans la liaison à la choline. Ces alignements ne permettent cependant pas de définir le nombre minimum de *repeats* nécessaires à la liaison. De même, nous ne possédons aucune information sur la façon dont le domaine de liaison se positionne sur le peptidoglycan ni sur le nombre de molécules de choline en contact avec un domaine de liaison.

Une approche permettant de poser des hypothèses consiste à trouver une structure connue relativement proche des domaines de liaison naturels. Pour ce faire, la séquence d'intérêt est soumise aux programmes de reconnaissance de *fold* 3D-PSSM (Kelley et al., 1999), UCLA.DOE (Fischer, 1999), PROSPECT V1.0 (Xu and Xu, 2000) et THREADER. (Jones et al., 1992).

Les critères de sélection que nous avons utilisés pour nos résultats sont les suivant :

1) **le score global pour une structure donnée**. De façon arbitraire, nous avons décidé d'analyser uniquement les dix premiers scores.

2) **la classification hiérarchique structurale de ces dix premiers hits**. Nous avons utilisé le programme de classification structurale des protéines CATH qui propose, pour toutes les structures connues, une classification en 4 niveaux (Orengo et al., 1997):

a) <u>la classe</u> : nature des structures secondaires majoritaires (1 = principalement des hélices α, 2 = principalement des brins β, 3 = un mélange d'hélices α et de brins β et 4 = peu de structures secondaires)

b) <u>le type d'architecture</u> : forme générale de la structure, déterminée par l'orientation des structures secondaires mais sans tenir compte des

liens entre les structure secondaires (60 = β sandwich, 10 = ribbon, 40 = barrel, etc...)

c) la topologie ou famille de *fold* : tient compte de la forme générale de la structure mais aussi de la façon et de l'ordre dans lequel les structures secondaires sont reliées les unes aux autres.

d) la superfamille des homologues : les structures sont regroupées en fonction de leur similarité structurale et de leur similarité de fonction.

En assignant un code de classification structural aux dix premières structures proposées, on peut avoir une première idée de la qualité des résultats obtenus. En effet, si un programme de reconnaissance de *fold* sélectionne, dans les premiers scores, des protéines de classe et d'architecture très différentes, les résultats doivent être interprétés très prudemment. A l'inverse, si les premiers scores proposent des protéines de même classe et même architecture, les résultats peuvent être considérés comme plus vraisemblables.

3) **la longueur de l'alignement séquence/structure**. Si un programme de reconnaissance de *fold* donne un très bon score pour un alignement séquence/structure mais que cet alignement ne reprend qu'une très petite portion de la séquence à laquelle on veut assigner une structure, la structure sélectionnée n'apportera que peu d'information par rapport à la structure du domaine complet. De même, un résultat de reconnaissance de *fold* fournissant un alignement sur toute la séquence soumise mais avec énormément d'insertions ou de délétions est sujet à caution car il signifie que l'on apportera beaucoup de modifications à la structure connue.

La comparaison de l'alignement séquence/structure au modèle topologique de la structure proposée permet de visualiser les modifications que l'alignement fourni apportera sur la structure.

4) lorsque les trois premiers critères sont rencontrés, il est possible de sélectionner des candidats à partir desquels on peut construire un modèle par comparaison (modélisation par comparaison ou *comparative modeling*). Cette technique implique l'assignation des coordonnées tri-dimentionnelles des atomes de la structure sélectionnée aux atomes des acides aminés composant le fragment de domaine (ou domaine entier) qui a été soumis pour la recherche. Dans ce cas, en plus des trois premiers critères propres à l'analyse des résultats de reconnaissance de *fold*, nous avons inclus dans notre méthode la recherche de une ou plusieurs cages de résidus aromatiques sur le modèle créé par homologie. Cette démarche supplémentaire repose sur des données de la littérature. En effet, l'acétylcholine estérase est une enzyme liant également la choline. L'étude structurale de son site actif a permis de définir les acides aminés impliqués dans la liaison à la choline (Harel et al., 1993). Il s'agit des résidus Tryptophane 84, Tryptophane 279 , Phénylalanine 330 et Sérine 200. Ces données renforcent les résultats obtenus lors de l'analyse en dichroïsme circulaire du domaine de liaison à la choline de la protéine LytA avec ou sans choline à savoir une forte contribution des résidus aromatiques dans la liaison à la choline (Usobiaga et

al., 1996) (Medrano et al., 1996). La définition de cette cage de résidus aromatiques étant la seule donnée expérimentale concrète que nous possédions sur la liaison d'une protéine à la choline, il nous a semblé important d'en tenir compte dans l'analyse de nos résultats au même titre, par exemple, qu'une portion de structure commune entre tous les modèles construits.

Les domaines complets de liaison à la choline ainsi que les fragments de domaine (2 ou 4 *repeats*) qui ont été sélectionnés pour cette analyse sont repris au tableau 38 (annexe 2). Cet échantillon présentait plusieurs caractéristiques :

- variation de la taille du domaine de liaison (2 à 15 *repeats*). En utilisant 2 *repeats*, nous espérions mettre en évidence des motifs structuraux courts pouvant éventuellement se répéter dans une structure complète. Le choix d'un fragment de quatre motifs se base sur des données expérimentales. En effet, comme nous l'avons mentionné dans l'introduction de ce travail, Garcia et ses collaborateurs ont montré que quatre motifs répétés du domaine de liaison C-LytA correspondaient expérimentalement au nombre minimum de motifs nécessaires pour la liaison à la choline (Garcia et al., 1994). De plus, les plus petits domaines de liaison à la choline existant dans la nature sont composés de 4 *repeats* et appartiennent aux protéines CbpG et PAL (Gosink et al., 2000) (Garcia et al., 1983). C'est donc à partir quatre *repeats* que nous envisageons de construire un modèle afin de définir les acides aminés importants pour la liaison.
- variation de la position des *repeats* testés au sein d'un domaine de liaison. En effet, lors de l'élaboration du consensus de séquences en classant préalablement les *repeats* en fonction de leur position dans le domaine de liaison à la choline, nous avions remarqué que le consensus établi sur base de l'alignement des *repeats* 1, c'est-à-dire des premiers *repeats* rencontrés à partir de l'extrémité N-terminale, différait, à certaines positions, des consensus établis sur base de l'alignement des motifs répétés aux autres positions. Nous en avions déduit que ce premier motif répété pouvait différer des autres parce qu'il servait de segment de jonction avec les autres domaines de la protéine. Dans ce contexte, il était important de ne pas utiliser systématiquement le premier *repeat* dans la recherche d'une structure homologue afin de ne pas biaiser le résultat.
- présence de domaines complets très homologues afin de tester si les différents programmes de reconnaissance de *fold* donnent des résultats concordants.
- protéines appartenant aux deux espèces bactériennes *Streptococcus pneumoniae* et *Clostridium beijerinckii* et à un phage de pneumocoque.

Dans le cadre de ce travail, rappelons que nous avons utilisé les serveurs de reconnaissance de repliements 3D-PSSM (Kelley et al., 1999), UCLA.DOE (Fischer, 1999), PROSPECT V1.0 (Xu and Xu, 2000) et THREADER (Jones, 1999).

Nous avons obtenus les résultats suivants :

- lorsque nous soumettons les séquences des domaines de liaison complets, les 5 candidats sélectionnés sont majoritairement composées de brin β et quatre d'entre-eux présentent une architecture de type sandwich β (tableau 39 annexe 2). Cependant, un essai de superposition des cinq structures ne permet pas de retrouver une portion de structure commune. Vu ce résultat, nous n'avons pas tenté de construire un modèle par homologie ni de rechercher des cages de résidus aromatiques au départ des cinq alignements séquence-structure proposés.
- lorsque nous soumettons les séquences de deux *repeats* consécutifs (incluant ou n'incluant pas le premier *repeat* N-terminal), un candidat préférentiel ressort (tableaux 40 et 41 annexe 2). Il s'agit du premier domaine de la fibronectine humaine (code PDB : 1fbr1). Cette protéine possède une structure en ruban β et présente 6 fois le motif brin β – coude – brin β dans son premier domaine, ce qui correspond à un motif structural répété. La construction d'un modèle par homologie au départ d'un des alignements séquence-structure proposé montre que les résidus aromatiques présents dans les deux *repeats* consécutifs ne s'organisent pas en cage.
- lorsque nous soumettons les séquences de quatre *repeats* consécutifs (incluant oui ou non le premier *repeat* N-terminal), nous obtenons au total 7 structures potentiellement proches de la structure d'un domaine de liaison à la choline (tableaux 42 et 43 annexe 2). Après construction d'un modèle par homologie au départ du meilleur alignement séquence-structure, 3 modèles sont susceptibles de présenter certains résidus aromatiques sous forme de cages. Cependant, nous constatons une grande variabilité dans l'ordre préférentiel des structures proposées. En effet, en fonction de la séquence de domaines répétés soumise au serveur de reconnaissance de *fold*, l'ordre préférentiel des structures proposées varie.

Les conclusions que nous pouvons tirer de ces différents essais sont les suivants :

- il suffit de changer légèrement la séquence des motifs répétés pour que de nouveaux candidats apparaissent. Cette variabilité dans les résultats ne nous permet pas de sélectionner une structure particulière qui pourrait être proche de la structure « *choline-binding* ».
- toutes les structures sélectionnées sont composées uniquement de brins β, ce qui renforce les prédictions de structures secondaires effectuées sur le motif consensus, à savoir deux brins β par motifs répétés.
- les topologies représentées sont soit le sandwich β soit le ruban β soit le *fold* « petite protéine ». Les sandwichs et rubans β présentent des motifs brin β-coude-brin β mais ces derniers ont rarement un arrangement répétitif dans l'espace. Or, le fait d'observer un nombre variable de motifs répétés dans les divers domaines de liaison à la choline nous laissait supposer que un ou plusieurs motifs répétés forment un motif structural précis permettant la liaison à la choline et que, en fonction du nombre total de *repeats* du

domaine considéré, ce motif structural se retrouvait un certain nombre de fois dans la structure. Nous nous attendions donc à retrouver parmi les résultats de reconnaissance de *fold*, des structures présentant un même arrangement structural répété plusieurs fois qui permettait ainsi d'expliquer l'existence intrinsèque des motifs répétés dans la séquence.

- la cage de résidus aromatiques trouvée sur deux modèles est composée de quatre acides aminés aromatiques et d'un résidu acide ou polaire.

En résumé, les expériences de reconnaissance de *fold* réalisées sur 2 *repeat*s consécutifs, 4 *repeat*s consécutifs ou des domaines complets ne permettent pas de sélectionner avec certitude une structure potentiellement proche des domaines de liaison à la choline.

Bien que les structures secondaires soient très vraisemblablement des brins β, nous ne pouvons prédire comment ces brins s'arrangent dans l'espace et ne pouvons donc définir quels sont les acides aminés du consensus indispensables pour la liaison à la choline.

I.4. Les structures des domaines de liaison à la choline ClytA et CPLI

A ce stade de notre travail, ne pouvant aller plus loin dans l'analyse bioinformatique des séquences des domaines de liaison à la choline, nous sommes passé à une phase expérimentale que nous détaillerons au chapitre suivant. Cependant, en décembre 2001, la structure cristallographique des *repeat*s 2 à 6 du domaine de liaison C-LytA, suivis de la queue C-terminale, a été publiée (Fernandez-Tornero et al., 2001). Par la suite, la structure du lysosyme du phage Cp-1 a été déterminée en octobre 2003 (Hermoso et al., 2003). Ce paragraphe est consacré à une description de ces structures ainsi qu'à l'analyse du consensus de séquence que nous avons établi en le confrontant aux structures.

I.4.1. Présentation générale de la structure du domaine de liaison partiel de l'amidase LytA de *Streptococcus pneumoniae*

Le monomère C-LytA a une forme générale cylindrique (Fernandez-Tornero et al., 2001). Les structures secondaires consistent en 6 épingles à cheveux (ou épingle) composées chacune de deux brins β anti-parallèles de même longueur (cinq résidus principalement hydrophobes), connectés par une courte boucle interne. Les épingles consécutives sont reliées par des boucles de 8 à 10 résidus (figure 17).
Les épingles s'étendent perpendiculairement à l'axe du cylindre, avec un pas de 120° entre-elles. Cette orientation des structures secondaires définit une superhélice gauche, les épingles i et i+3 se superposant. Le squelette de la structure du domaine partiel de liaison à la choline C-LytA peut donc être considéré comme un escalier en spirale, avec trois marches par tour.

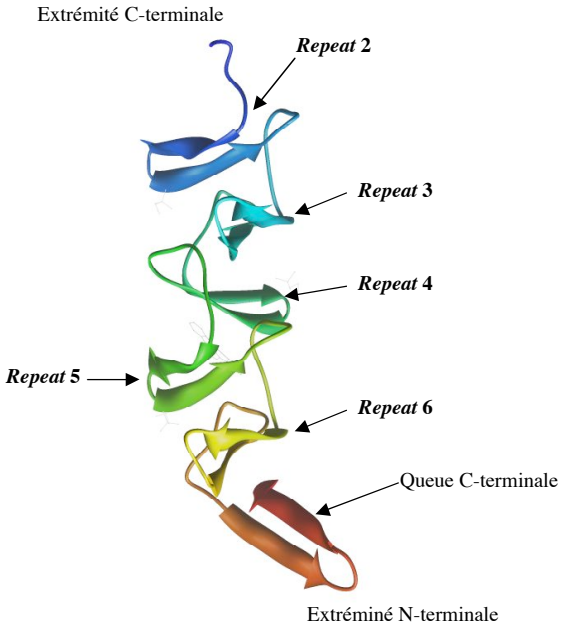


Figure 17 Présentation de la structure du domaine de liaison à la choline de l'amidase LytA. Les flèches symbolisent les brins β (résolution de la structure : 2,6 Å - code PDB : 1hcx).

En se basant sur cette topologie, on peut considérer que le domaine de liaison à la choline de l'amidase LytA appartient à la famille structurale des solénoïdes. Les structures solénoïdes présentent un arrangement en superhélice d'unités structurales répétées. Dans le cas de la protéine LytA, l'unité structurale de base est une épingle à cheveux constituée de deux brins β de même longueur, précédée d'un boucle de 8 à 10 résidus (figure 18).

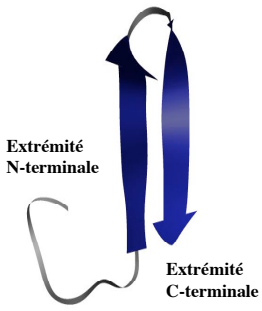


Figure 18 Illustration d'une unité structurale correspondant à un *repeat* de séquence du domaine de liaison à la choline de l'amidase LytA. Chaque unité structurale est composée d'une boucle suivie d'une épingle à cheveux.

Cependant, malgré les similarités partagées avec les solénoïdes-β déjà répertoriés, le domaine de liaison à la choline de l'amidase LytA diffère de ces derniers car il est construit au départ de superstructures secondaires (les épingles), possédant leur propre entité. On peut donc considérer que la structure de C-LytA représente **un nouveau *fold* protéique (*left-handed ββ-3-solenoid spiral staircase*)**. Cette originalité justifie pourquoi nous n'avons pas pu sélectionner la structure d'une protéine similaire par les méthodes de reconnaissance de *folds*.

Si on compare la séquences des *repeats* 2 à 6 ainsi que celle de la queue C-terminale à la structure, on peut relever les caractéristiques suivantes :

- les *repeats* de séquence correspondent bien à des *repeats* structuraux avec deux unités structurales distinctes par *repeats* : une épingle β précédée d'une boucle de 8 à 10 résidus,
- les résidus aromatiques présents dans le deuxième brin β de l'épingle sont strictement conservés alors que le premier brin présente plus de variabilité,
- les boucles de 8 à 10 résidus présentent également moins d'identité de séquence.

Une analyse plus poussée de la structure montre que la queue C-terminale se distingue des *repeats*. Bien qu'elle forme une épingle précédée d'une boucle, comme les autres *repeats*, l'angle formé entre l'épingle du *repeat* 6 et celle de la queue C-term n'est que de 95° au lieu de 120°. De plus, cette dernière épingle n'est pas exactement perpendiculaire à l'axe du cylindre. Ces caractéristiques structurales particulières correspondent au peu de

similarité de séquence observée entre la queue C-terminale et les autres *repeats*.
En solution, les domaines de liaison C-LytA forment naturellement des dimères. La forme générale du dimère ressemble à un boomerang, portant probablement aux extrémités de ces bras les domaines catalytiques N-terminaux (figure 19). L'interaction de deux monomères semble se faire principalement par appariement des deux épingles portées par les repeats 6 et les deux queues C-terminales(Fernandez-Tornero et al., 2002a). Dans toutes les épingles conservées, les résidus hydrophobes et aromatiques sont orientés vers l'intérieur, formant un noyau hydrophobe qui constitue la principale force de dimérisation.

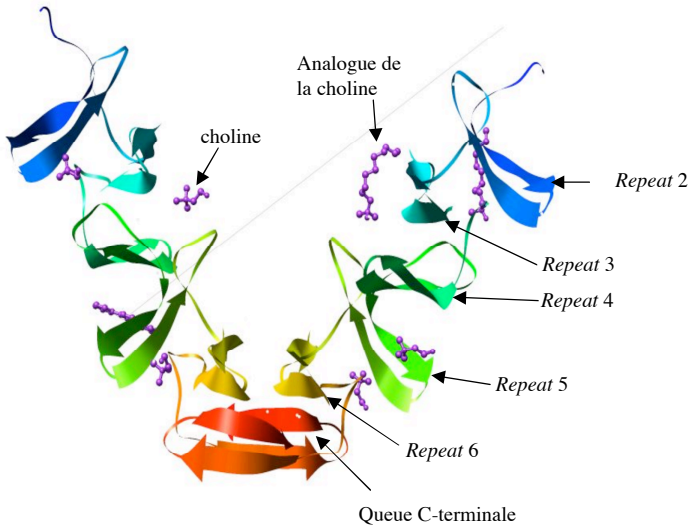


Figure 19 Présentation du dimère de C-LytA.

Sur la structure, quatre sites de liaison à la choline sont répertoriés. Ces sites correspondent à des cages de résidus aromatiques. Ils sont formés par l'interface de deux épingles consécutives : épingles 2 et 3 pour le premier site, épingle 3 et 4 pour le deuxième site, épingles 4 et 5 pour le troisième site et épingles 5 et 6 pour le quatrième site (tableau 13).

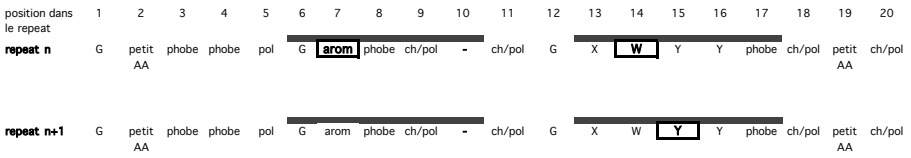


Tableau 13 Localisation des résidus aromatiques contribuant à la formation d'un site de liaison à la choline. Ces résidus sont surlignés en gras et encadrés
Pol : classe des résidus polaires
G : glycine
Arom : classe des résidus aromatiques
Phobe : classe des résidus hydrophobes
Ch/pol : classes des résidus chargés ou polaires
X : position variable

La nature de l'interaction est principalement hydrophobe avec les trois groupements méthyle de la choline remplissant une cavité étroite de 15 $Å^3$ formée par trois résidus aromatiques et un résidu hydrophobe. Une interaction cation-π entre les sytèmes riches en électrons des anneaux aromatiques et la charge positive de la choline renforce la liaison (figure 20).

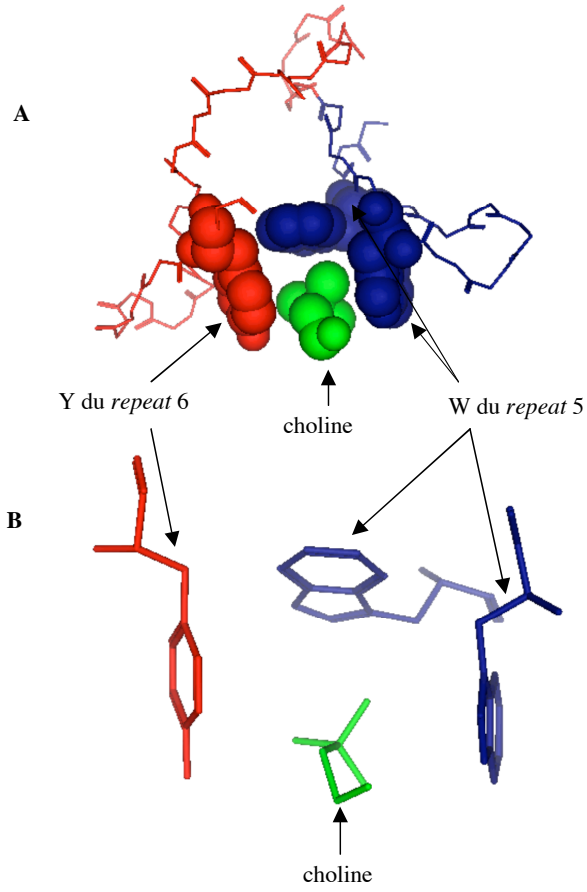


Figure 20 Présentation d'un site de liaison à la choline.
A : représentation sous forme "*space filling*" d'une cage de résidus aromatique servant de site de liaison à une molécule de choline.
B : représentation de la même cage de résidus aromatiques et de la molécule de choline.

Comme d'autres hélices, la structure du monomère C-LytA présente en surface des sillons en spirale, reliant les différents sites de liaison entre-eux. Formés par des résidus chargés et polaires, ces sillons ont une longueur de 10 Å, une profondeur de 7 Å et une ouverture maximale de 15 Å. Or, rappelons que, selon la structure de l'acide lipotéichoique définie par spectroscopie RMN, ce dernier est composé de 2 à 8 unités répétées glucidiques, portant chacune deux molécules de phosphocholine. La distance entre les têtes hydrophobes des phosphocholines peut être réduite, par simple torsion, à 10 Å. On peut donc penser que le sillon entre deux sites de liaison consécutifs correspond à une unité de glucan des acides téichoique et lipotéichoique de la paroi du pneumocoque. De plus, les atomes d'oxygène et d'azote des résidus N-acétylgalactosamine des unités

de glucans pourraient établir des ponts hydrogènes et des interactions électrostatiques avec des atomes polaires correspondant, formant les sillons dans la structure du domaine de liaison C-LytA. Ces données permettraient d'expliquer la grande affinité du domaine C-LytA pour la paroi du pneumocoque.
Un essai de cristallisation du domaine complet de liaison C-LytA a été réalisé (Fernandez-Tornero et al., 2002b). Outre les *repeats* 2 à 6 et la queue C-terminale déjà mentionnés ci-dessus, le domaine C-LytA cristallisé possède 19 résidus additionnels en N-terminal, correspondant au *repeat* 1. Bien que les cartes de densité électronique ne permettent pas de positionner avec précision un grand nombre de chaînes latérales et qu'il y ait des discontinuité à des endroits spécifiques, les résultats suggèrent que le premier *repeat* adopte la même conformation que les autres *repeats* et qu'il y aurait donc formation d'un site de liaison à la choline supplémentaire entre les épingles des *repeats* 1 et 2.

I.4.2. Présentation générale de la structure du lysosyme CPL1 du phage Cp-1

La chaîne polypeptidique du lysosyme CPL1 consiste en un domaine catalytique relié au domaine de liaison à la choline par un connecteur composé de dix résidus acides (figure 21) (Hermoso et al., 2003).

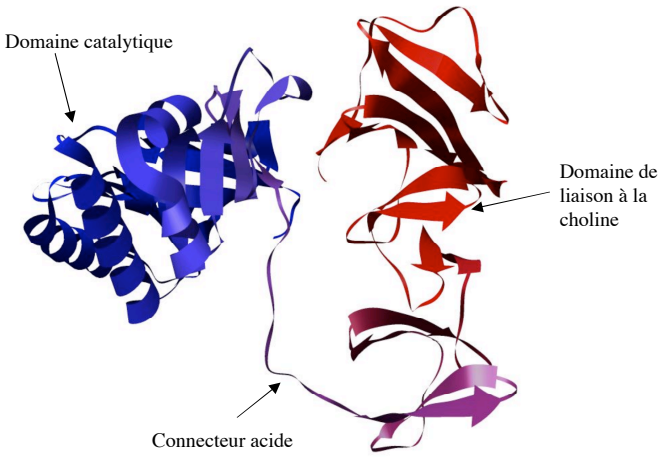


Figure 21 Structure générale du lysosyme CPL1. (résolution de la structure : 2,45 Å- code PDB : 1ho9).

Comme pour l'amidase LytA, le domaine de liaison est constitué de 6 *repeats* d'une vingtaine d'acides aminés, adoptant chacun comme structure

secondaire, une épingle β suivie d'une boucle. Par contre, ces *repeats* sont répartis dans deux régions structurales bien distinctes, appelées CI et CII (figure 22). Le domaine structural CI est constitué des quatre premiers *repeats* adoptant le même arrangement en superhélice gauche que les *repeats* de LytA. Les *repeats* 5 , 6 et la queue C-terminale, constituant le domaine structurale CII, adoptent une structure en feuillet βavec 6 brins β anti-parallèles. Le domaine de liaison à la choline du lysosyme CPL1 consiste donc en un arrangement nouveau et distinct de 6 séquences répétées.

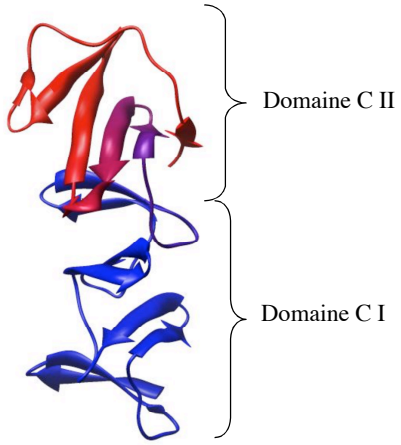


Figure 22 Illustration des deux régions structurales du domaine de liaison à la choline du lysosyme CPL1. L'arrangement en superhélice gauche des quatre premiers *repeats* constitue de domaine CI tandis que le feuillet β composé des deux derniers *repeats* et de la queue C-terminale constitue le domaine CII

Ces différences observées entre les deux protéines pourraient provenir soit de différences dans la séquence des *repeats* soit des interactions entre les modules catalytique et de liaison du lysosyme CPL1. La deuxième hypothèse semble la plus probable. En effet, l'étude structurale de CPL1 montre que l'interface entre les modules catalytique et de liaison est construite par une cavité hydrophobe comprise entre les brins β 6 et β 8 du domaine catalytique, où la queue C-terminale du domaine de liaison vient s'insérer, protégeant ainsi l'interface hydrophobe du solvant. Selon les auteurs, ces contraintes intermodulaires permettraient d'orienter de façon optimale le domaine catalytique dans le substrat de polysaccharides.

Sur base de la structure de l'amidase LytA, quatre sites de liaison à la choline sont prédits pour le lysosyme. Cependant, la structure de CPL1 montre que, si la géométrie de ces sites est conservée (formation de cages de résidus aromatiques), seuls les deux premiers sites semblent fonctionnels. Les deux derniers sites sont masqués par des résidus suite à la disruption de la superhélice pour former le feuillet β.
Dans le cas du lysosyme, un site de liaison consiste en une cage de trois résidus aromatiques et d'une lysine venant coiffer le site et pouvant stabiliser par un pont hydrogène le groupement phosphate de la phosphoryl-choline.

En résumé, les deux structures présentées ci-dessus confirment nos analyses bioinformatiques puisque les structures secondaires sont bien constituées de brins β et que le *fold* adopté par les domaines de liaison à la choline n'a pas d'équivalent dans les banques de données. Ils nous apprennent également que les *repeat*s constituant les domaines de liaison à la choline peuvent adopter deux types de conformation :
- soit un *repeat* se replie en une épingle β suivie d'un boucle. Les épingles ainsi formées s'organisent ensuite en une superhélice gauche,
- soit les *repeat*s adoptent une conformation en feuillet β.

Lorsque les *repeat*s s'organisent en superhélice gauche, il y a formation de sites de liaison à la choline.
Un site de liaison consiste en une cage de trois résidus aromatiques répartis sur deux *repeat*s consécutifs.

I.4.3. Analyse du consensus de séquences établi par confrontation avec les structures cristallographiques de ClytA et CPL1

Une première comparaison de notre consensus aux deux structures nous permet de dégager les observations suivantes (tableau 14).
Les structures secondaires observées expérimentalement sont bien constituées de brin β, de coudes et de boucles, comme prédit. Cependant, dans la structure de l'amidase LytA, les brins β ont tous la même longueur (5 résidus). Dans la structure du lysosyme CPL1, cette longueur varie de 4 à 6 résidus et dans le consensus de structures secondaires prédites, un brin était constitué de 7 acides aminés et l'autre de 4 acides aminés.
Selon l'auteur considéré, un *repeat* de structure débute soit par la boucle, soit par l'épingle β. Dans la suite de notre travail, nous avons décidé, de façon arbitraire, de définir un *repeat* structural comme étant constitué d'une épingle β suivie d'une boucle Le premier résidu d'un *repeat* est donc un résidu polaire.

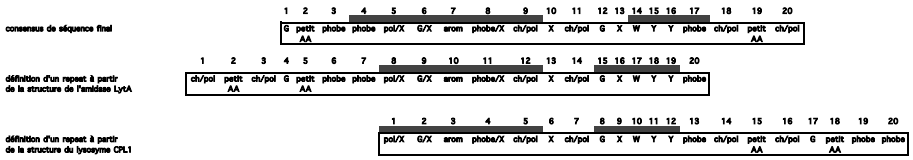


Tableau 14 Comparaison du consensus de *repeat* (paragraphe I.2.4.3) au *repeat* défini au départ de la structure partielle du domaine de liaison à la choline C-LytA et de la structure du domaine de liaison du lysosyme CPL1.

Afin d'analyser et, si possible, de valider les classes d'acides aminés que nous avions jugées importantes à diverses positions du consensus, nous avons analysé les contacts hydrophobes, les formations de ponts hydrogènes et de ponts salins qui existent entre les résidus des structures. Ces données ont été calculées par le logiciel de modélisation moléculaire MOE (Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada). Les seuils de distance utilisés pour définir ces interactions sont repris au tableau 44 de l'annexe 3. Puis, nous avons sélectionné les séquences de quatre *repeats* consécutifs et ce, pour trois domaines de liaison différents (tableau 45 annexe 3). Les *repeats* de ces domaines ont été choisis car les résidus les composant répondaient particulièrement bien au consensus général établi. Nous avons modélisé ces trois fragments de domaine en utilisant soit les coordonnées tri-dimentionnelles des *repeats* 2, 3, 4 et 5 de la structure C-LytA, soit les coordonnées tri-dimentionnelle des *repeats* 1, 2, 3 et 4 du domaine de liaison CPL1.
Ces modélisations ont été réalisées en fixant tout l'alignement séquence - structure, c'est-à-dire en obligeant le programme de modélisation du logiciel MOE à transposer tel quel les résidus de la séquence sur les résidus de la structure, ne laissant des degrés de liberté que pour le choix des rotamères (orientation des chaînes latérales). Comme pour l'analyse des structures, le but de ces modélisations est de visualiser dans l'espace la position de chaque classe de résidus par rapport à la structure globale et d'évaluer les contacts possibles de cette classe de résidus avec les autres acides aminés et sa contribution éventuelle à la stabilité de la structure. Seuls ont été retenus les contacts retrouvés systématiquement dans au moins 4 des 6 modèles réalisés. Les positions dans le consensus ont été définies au départ de la structure de CLytA.
Nous obtenons les résultats suivants :

Position 1 : classe des résidus polaires - X avec prédominance de thréonines: ces résidus sont situés dans le coude juste avant l'épingle β, leur chaîne latérale pointant vers l'extérieur. On peut supposer que ces résidus intéragissent avec les acides téichoique et lipotéichoique constituant la paroi bactérienne.

Position 2 : classe des résidus glycine - X avec prédominance de glycines. Ces résidus sont situés dans le coude avant l'épingle β. La prépondérance des glycines à cette position semble logique puisqu'il s'agit d'un résidu formateur de coude possédant une très petite chaîne latérale et permettant ainsi d'éviter ainsi les encombrements stériques.
Position 3 : classe des résidus aromatiques avec prépondérance de tryptophanes. Ces résidus sont situés dans le premier brin β de l'épingle. Il s'agit d'un des trois résidus formateurs des sites de liaison (tableau 12). Les aromatiques situés à cette position réalisent de nombreux contacts hydrophobes, notamment avec les aromatiques en position 10 (également formateurs des sites de liaison) et les résidus hydrophobes situés en positions 13 et 19 du même *repeat*. Ils ont également des contacts hydrophobes avec le résidu hydrophobe situé en position 19 du *repeat* n+1.
Position 4 : classe des résidus hydrophobes - X : ces résidus appartiennent au premier brin β de l'épingle β. Ils établissent soit des interactions hydrophobes soit des ponts hydrogènes en fonction de leur nature et des autres résidus présents dans la séquence du même *repeat*. La nature des résidus à cette position est donc assez variable.
Position 5 : classe des résidus chargés - polaires. Ils se situent dans le premier brin β. Leur chaîne latérale est orientée totalement vers l'extérieur. Si nous n'avons pu définir des interactions particulières avec d'autres résidus de la structure, il est possible que ces résidus contribuent, comme ceux de la position 1, aux interactions avec la paroi bactérienne.
Position 6 : classe des résidus X, c'est-à-dire tout à fait variable en fonction du domaine de liaison considéré. Nous n'avons pu établir aucune interaction privilégiée de ce résidu avec le reste de la structure, probablement de par sa nature variable. Ces résidus constituent le dernier acide aminé du premier brin β.
Position 7 : classe des résidus chargés - polaires. Ces acides aminés sont situés dans le coude entre les deux brins β. Ils n'établissent aucun contact particulier avec d'autres résidus de la structure.
Position 8 : classe des glycines. La présence d'une glycine comme deuxième résidu d'un coude peut s'expliquer par le faible volume de leur chaîne latérale permettant d'éviter les encombrements stériques.
Position 9 : classe des résidus X. Ces résidus constituent le premier résidu du deuxième brin de l'épingle β. Aucun contact préférentiel n'a été mis en évidence vu la nature variable des acides aminés situés à cette position.
Position 10 : classe des résidus aromatiques avec forte prépondérance des tryptophanes. Ces derniers, situés dans le deuxième brin, participent à la formation des sites de liaison. Ils réalisent principalement des contacts hydrophobes avec les résidus situés en position 3 et 19 du même *repeat* mais également avec le résidu en position 19 du *repeat* suivant.
Position 11 : classe des résidus aromatiques avec forte prépondérance des tyrosines. Toujours situés dans le deuxième brin β, ces résidus participent à la formation des sites de liaison. Mise à part la formation ponctuelle de ponts

hydrogène dans la structure C-LytA (et pas pour toutes les tyrosines situées en position 11 des cinq *repeats* considérés), ce résidu ne semble pas établir d'interactions systématiques avec d'autres acides aminés du domaine. Sa chaîne latérale est cependant toujours orientée vers l'extérieur. On peut donc à nouveau suggérer une interaction de l'hydroxyle de ce résidu avec les acides téichoïques et lipotéichoïques constituant la paroi.
Position 12 : classe des résidus aromatiques avec forte prépondérance des tyrosines. Dans les structures et modèles établis, nous n'avons pas pu mettre en évidence d'interaction systématique avec d'autres résidus du domaine. Cependant, nous pouvons quand même relever dans les deux structures, l'établissement sporadique de contacts hydrophobes avec l'aromatique situé en position 3 (formateur de cage) du même *repeat* ainsi qu'avec les résidus n°6 (X) et n°13 (hydrophobe) du *repeat* n+1. De même, nous observons de façon non systématique l'établissement d'un pont hydrogène entre le groupement hydroxyle de la chaîne latérale de la tyrosine et un autre résidu de la structure .
Position 13 : classe des résidus hydrophobes avec prépondérance de leucines. Derniers résidus du deuxième brin β de l'épingle, ces acides aminés établissent des contacts hydrophobes avec les résidus hydrophobes situés en position 19 du même *repeat*.
Position 14 : classe des résidus chargés - polaires. Situés juste après l'épingle, ces acides aminés semblent réaliser des ponts hydrogène ou salins avec d'autres résidus de la structure mais pas de façon systématique. Leur chaîne latérale est orientée vers le solvant, ce qui peut suggérer une interaction avec la paroi bactérienne.
Position 15 : présence prépondérante de petits résidus. Ceux-ci se situent dans le coude après l'épingle β. Il est donc relativement logique de retrouver à cette position une prépondérance de résidus possédant une chaîne latérale peu volumineuse, évitant ainsi les encombrements stériques. Leur chaîne latérale est également orientée vers le solvant, pouvant indiquer une interaction avec les acides téichoïque et lipotéichoïque de la paroi.
Position 16 : classe des résidus chargés - polaires. Ces résidus constituent le deuxième acide aminé du coude. Aucune interaction spécifique avec d'autres résidus de la structure n'a pu être mise en évidence.
Position 17 : classe des glycines. Elles sont situées juste après le coude au début de la boucle. Il est probable que leur présence à cette position soit liée à un problème d'encombrement stérique.
Position 18 : présence prépondérante de petits résidus. Sur base des modèles créés, nous ne pouvons justifier l'importance de ces résidus à cette position, si ce n'est un problème d'encombrement stérique.
Position 19 : classe des résidus hydrophobes. Situés dans la boucle, ces acides aminés établissent des contacts hydrophobes systématiques avec les résidus hydrophobes situés en position 13 du même *repeat*. De façon moins systématique, on relève également des contacts hydrophobes avec les résidus situés en position 10 du *repeat* n-1.

Position 20 : classe des résidus hydrophobes. Situés dans la boucle, ces résidus ne semblent pas développer d'interactions systématiques avec d'autres résidus.

En conclusion, de manière générale, les acides aminés chargés - polaires aux positions 1, 5, 7, 14, 16 et le résidu hydrophobe en position 20 interagissent peu avec d'autres résidus des deux structures. Cependant, leur chaîne latérale est le plus souvent orientée vers l'extérieur, ce qui pourrait suggérer une interaction avec les atomes d'oxygène et d'azote ou les groupements méthyles des sucres des unités de glucans. Cette hypothèse est en concordance avec les interprétations de la structure du domaine de liaison ClytA (Fernandez-Tornero et al., 2001).

Les glycines et petits résidus aux positions 2, 8, 15, 17 et 18 sont situés dans des zones de torsion de la structure où d'éventuels problèmes d'encombrement stérique pourraient survenir.

Les résidus aromatiques situés en position 3, 10 du *repeat* n et à la position 11 du *repeat* n+1 sont les trois acides aminés formant le site de liaison à la choline. Il est donc logique qu'ils soient très conservés. Outre ces trois résidus, l'analyse du consensus révèle l'importance des résidus hydrophobes situés aux positions 13 et 19. D'après nos analyses, le résidu 19 (le plus souvent une méthionine) semble constituer le plancher des cages d'aromatiques. De son côté, le résidu 13 semble systématiquement en contact avec le résidu 19, établissant probablement une liaison hydrophobe et contribuant de ce fait à stabiliser la structure. Dans les articles décrivant les structures C-LytA et CPL1, ces résidus sont considérés comme conservés à 50 - 60%. Cependant, aucun rôle ne leur avait été attribué.

Les classes situées aux positions 4, 6 et 9 étant trop variables d'un domaine à l'autre, aucun rôle structural général n'a pu leur être attribué. Une nouvelle analyse des 126 *repeats* à ces positions ne nous a pas permis de mettre en évidence des couples préférentiels (hydrophobe–hydrophobe, positif–négatif, polaire–polaire, polaire–chargé, chargé–polaire) à ces positions.

I.5. Conclusions de l'analyse bioinformatique

Dans le cadre de la création d'un *tag* de purification au départ du domaine de liaison à la choline de l'amidase LytA, nous avons d'abord développé une approche bioinformatique. Notre première étape a consisté à rechercher tous les domaines de liaison à la choline décrits dans les banques de séquences.

Ces domaines étant constitués de séquences répétées d'une vingtaine de résidus, nous avons, ensuite, étudié les caractéristiques physico-chimiques des séquences complètes des domaines de liaison à la choline. Aucune périodicité particulière dans la distribution des résidus n'a pu être mise en évidence, si ce n'est la présence d'un triplet de résidus aromatiques déjà décrit dans la littérature.

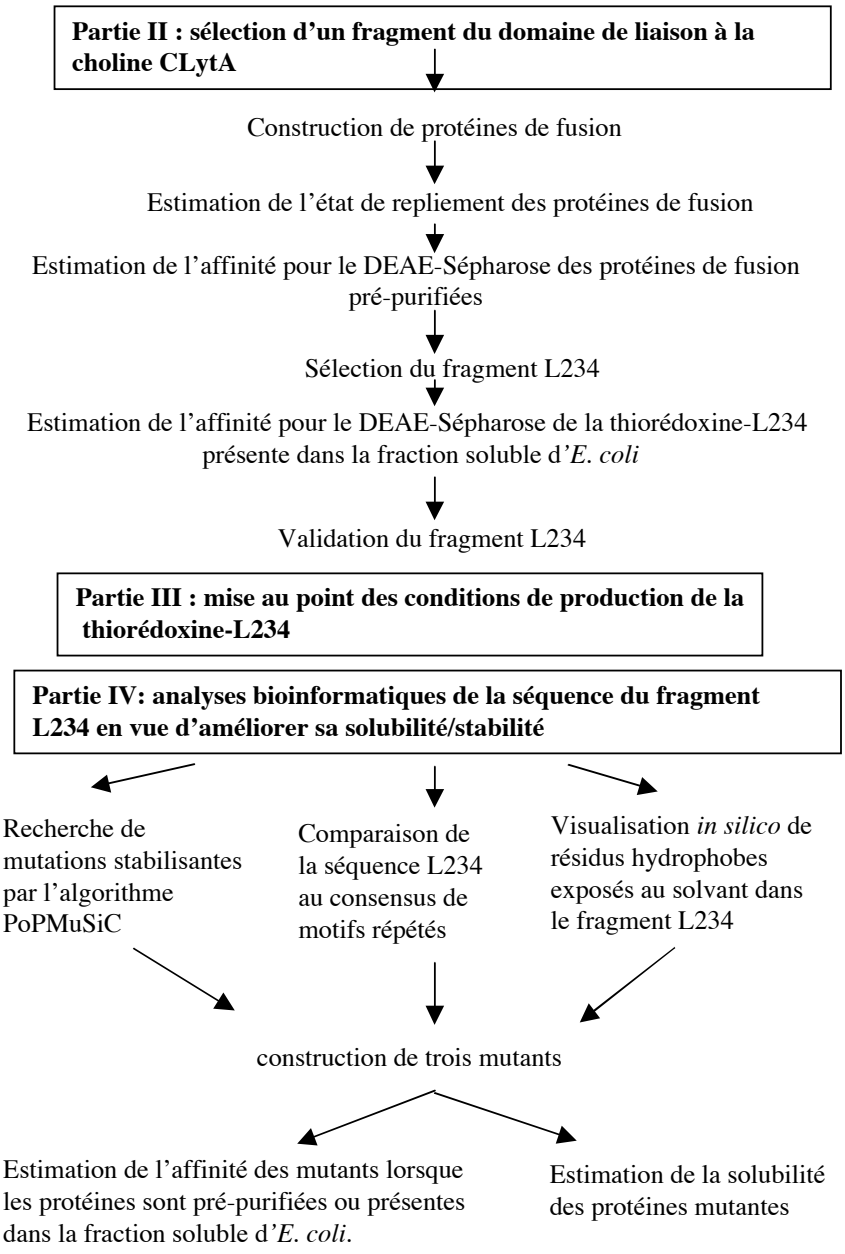
L'étape suivante a consisté à aligner les séquences répétées des domaines de liaison sélectionnés par divers programmes d'alignements. Nous avons ainsi pu établir un consensus de séquences au départ de 126 *repeats*, ce qui n'avait jamais été fait dans la littérature.

En parallèle, nous avons recherché dans la banque de structures PDB une structure potentiellement proche du domaine de liaison C-LytA. En recherchant dans la banque de données NR/PDB à l'aide du logiciel PSI-BLAST (Altschul et al., 1997), nous n'avons pas pu sélectionner une protéine de structure connue dont la séquence présentait un grand pourcentage de similarité. La méthode de modélisation par homologie n'était donc pas possible. Nous nous sommes alors tournés vers les méthodes de reconnaissance de *fold* qui permettent, dans certains cas, de sélectionner des protéines possédant la même structure générale tout en présentant un faible pourcentage de similarité de séquence. De nouveau, cette approche ne nous a pas permis de sélectionner une structure potentiellement proche du domaine de liaison à la choline C-LytA.

Lors de la parution de la structure partielle du domaine de liaison de l'amidase LytA en 2001, il s'est avéré que ce domaine de liaison présentait un nouveau *fold*, non encore répertorié dans les banques de données. Il est donc logique que la modélisation par homologie et la recherche d'une structure potentiellement proche par méthode de reconnaissance de *fold* n'aient pas abouti. L'analyse de cette structure, ainsi que celle du domaine de liaison CPL1 parue en 2003, nous a permis d'étudier notre consensus d'un point de vue structural. Notre conclusion est qu'une partie des classes de résidus que nous savions définies comme conservées à diverses positions d'un *repeat* type remplit bien soit un rôle structural (zones de torsion dans la structure) soit un rôle dans la formation des sites de liaison à la choline. Pour d'autres positions, nous pouvons suggérer un rôle d'interaction avec la paroi bactérienne. Nous n'avons pu attribuer de fonction pour trois positions considérées comme hyper-variables dans notre consensus. Il faudrait la parution d'autres structures de domaines de liaison à la choline pour analyser plus en profondeur ces positions.

Une seconde phase de l'approche bioinformatique concerne la sélection de mutations stabilisantes par l'algorithme PoPMuSiC (Gilis and Rooman, 2000). Nous la décrirons dans le cadre des résultats expérimentaux.

Représentation schématique des trois étapes de l'approche expérimentale.



PARTIE II : SELECTION D'UN FRAGMENT DU DOMAINE DE LIAISON CLYTA

L'analyse bioinformatique développée dans la première partie des résultats ne nous a pas permis de sélectionner une structure relativement proche de celle du domaine de liaison ClytA. Nous n'avons donc pas pu définir les éléments structuraux importants conférant au domaine de liaison ClytA son affinité pour la choline et pour le DEAE. Si l'établissement d'un consensus de séquence a permis de mettre en évidence certains résidus ou classes de résidus conservés, l'information n'est cependant pas suffisante pour définir la séquence d'un *tag* à partir de ces seuls résultats.
Une autre approche consiste à sélectionner expérimentalement un petit fragment de domaine présentant une affinité spécifique pour le DEAE-Sépharose. Ce travail constituera la deuxième partie des résultats. Ensuite, en fonction du rendement protéique obtenu pour la protéine de fusion sélectionnée, nous nous attacherons à définir des conditions de production optimales (troisième partie) et à définir, par diverses approches bioinformatiques, des mutations ponctuelles pouvant potentiellement améliorer les propriétés physico-chimiques et d'affinité du *tag* (quatrième partie). Les trois étapes expérimentales développées dans notre travail sont synthétisées dans l'organigramme présenté à la page précédente.

II.1. Dissection du domaine de liaison ClytA

II.1.1. Construction des protéines de fusion thiorédoxine-domaine de liaison ClytA tronqué

La première étape de l'approche expérimentale a consisté à sélectionner le plus petit fragment de domaine CLytA présentant toujours une affinité spécifique pour le DEAE.
Dans le cadre de ce travail, nous avons défini qu'un fragment protéique possédait une affinité spécifique pour le DEAE à partir du moment où il nécessitait l'ajout de choline (l'analogue structural du DEAE) pour s'éluer du support DEAE par compétition. Les polypeptides s'éluant en présence d'un sel tel que le NaCl ont été considérés comme non affins puisque leur adsorption sur le support DEAE résultait plutôt d'interactions électrostatiques non spécifiques.
Cette étape du travail se justifie par le peu de données et l'hétérogénéité des résultats présents dans la littérature. En effet, des expériences de délétions progressives au départ de l'extrémité C-terminale du domaine de liaison ClytA avaient montré que l'amidase LytA, présente dans un extrait protéique brut, devait posséder au moins les 4 premiers *repeat*s de son domaine de

liaison pour conserver une affinité spécifique pour des filtres de DEAE-cellulose (Garcia et al., 1994), la protéine L123 ne possédant plus que les trois premiers *repeats* s'éluant à 0,75M NaCl. Le domaine de liaison à la choline de la protéine PspA a également été étudié (Yother et al., 1992) (Yother and White, 1994). Ces auteurs ont montré que, sur les dix *repeat*s initialement présents dans le domaine, cinq sont nécessaires pour garder une liaison spécifique sur une colonne DEAE-cellulose lorsque la protéine est purifiée au départ d'un surnageant de culture. Les cinq premiers motifs répétés du domaine de liaison sont également indispensables pour que la protéine PspA reste adsorbée sur la paroi de la bactérie *in vivo*.
Le nombre de *repeat*s et les supports utilisés variant, il nous a semblé opportun de redéfinir le nombre minimum de *repeat*s nécessaires pour conserver une affinité spécifique pour le DEAE lorsqu'on utilise comme support du DEAE-Sépharose Fast Flow. De plus, seules des délétions progressives au départ de l'extrémité carboxy-terminale avaient été réalisées, ne donnant aucune information sur l'affinité de *repeat*s carboxy-terminaux en absence de motifs répétés N-terminaux.
Afin de déterminer le plus petit fragment de domaine possédant toujours une affinité spécifique pour le DEAE-Sépharose, nous avons donc créé huit protéines de fusion (figure 23 et tableau 46, annexe 4). Chaque protéine de fusion est constituée de la protéine reporter thiorédoxine, étiquetée en N-terminal par un *tag* 6 Histidines et fusionnée en C-terminal à un fragment de domaine LytA variant dans sa composition en *repeat*s.
Chez *E. coli*, la thiorédoxine est une petite protéine très soluble (14857 Da). Lorsqu'elle est surexprimée dans *E. coli* K12, elle peut représenter jusqu'à 40% des protéines cellulaires totales et, même à de tels taux d'expression, la majorité de la protéine produite reste sous forme soluble (Lunn et al., 1984). Cette caractéristique lui vaut d'être utilisée en routine dans des systèmes d'expression en tant que partenaire de fusion (laVallie et al., 1993). Elle constitue donc un candidat de choix comme protéine reporter pour la construction de nos protéines de fusion.
En pratique, chaque fragment de domaine a été amplifié par PCR au départ du domaine complet et cloné en aval de la thiorédoxine aux sites de restriction *Asp*718 et *Sac* I dans le vecteur pET15b2 (matériel et méthodes, paragraphe 2).
Après séquençage des constructions, les protéines ont été surexprimées dans les bactéries BL21(λDE3) et purifiées sur colonne de chélation (matériel et méthode, paragraphe 5). Bien que les constructions aient été confirmées par séquençage, la protéine de fusion thiorédoxine-L345 n'est pas produite tandis que la quantité de thiorédoxine-L456 soluble est très faible, la majorité de cette protéine précipitant sous forme de corps d'inclusion. Il est possible que ces protéines de fusion soient instables et se fassent dégrader au fur et à mesure qu'elles sont produites. Pour la thiorédoxine-L345, une autre hypothèse est la toxicité de cette protéine de fusion pour la bactérie. Ce candidat a été écarté pour la suite de nos analyses.

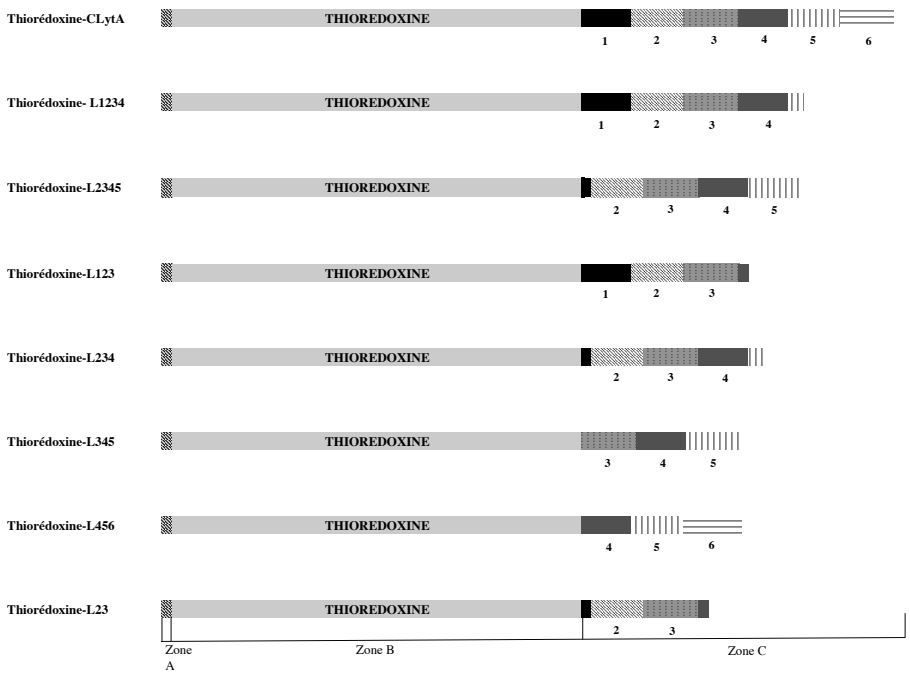


Figure 23 Représentation schématique des protéines de fusion thiorédoxine-fragment de domaine de liaison à la choline
Zone a : *tag* 6 Histidines
Zone b : protéine reporter thiorédoxine
Zone c : fragment du domaine de liaison à la choline

II.1.2. Estimation indirecte de l'état de repliement des protéines de fusion

L'impossibilité de produire les protéines thiorédoxine-L345 pose la question de l'état de repliement et, de façon indirecte, de la stabilité des protéines de fusion construites. En effet, il est possible que l'ajout d'un fragment du domaine CLytA en C-terminal de la thiorédoxine empêche la protéine de fusion d'adopter une structure stable, provoquant la dégradation de la protéine de fusion au fur et à mesure où elle est produite. Ce phénomène expliquerait en partie les faibles quantités de protéines solubles obtenues de manière générale pour les protéines de fusion (14 mg/l pour la thiorédoxine-CLytA contre 50 mg/l pour la thiorédoxine). Dans cette optique, nous avons donc mesuré la fluorescence intrinsèque des protéines de fusion produites dans *Escherichia coli* en présence de concentrations croissantes en chlorure de guanidium (matériel et méthodes, paragraphe 11). Lorsque la concentration en agent dénaturant augmente, un changement de l'intensité de fluorescence émise par les résidus aromatiques témoigne d'un changement dans leur environnement et, de ce fait, un changement de structure de la

protéine. Donc, si l'on observe un changement de l'intensité de fluorescence intrinsèque d'une protéine de fusion thiorédoxine-domaine de liaison tronqué CLytA lorsqu'on ajoute des concentrations croissantes en chlorure de guanidium, on peut en déduire que cette protéine est probablement en train de se dénaturer et qu'elle avait donc adopté une certaine structure lors de sa production dans la bactérie.
En pratique, 200 µg de protéines, resuspendues dans un volume final de 1 ml de tampon PBS pH 7,9 ont été incubés 30 minutes à température ambiante avec du chlorure de guanidium 0M, 1M, 2M, 3M, 4M, 5M, 6M, 7M et 8M. La fluorescence de chaque échantillon a ensuite été mesurée à 340 nm (longeur d'onde d'émission), en utilisant une longueur d'onde d'excitation de 282 nm. La figure 24 présente les résultats obtenus pour les protéines thiorédoxine-ClytA, thiorédoxine-L456 et thiorédoxine-L234 tandis que le tableau 15 résume nos observations pour toutes les protéines de fusion testées.

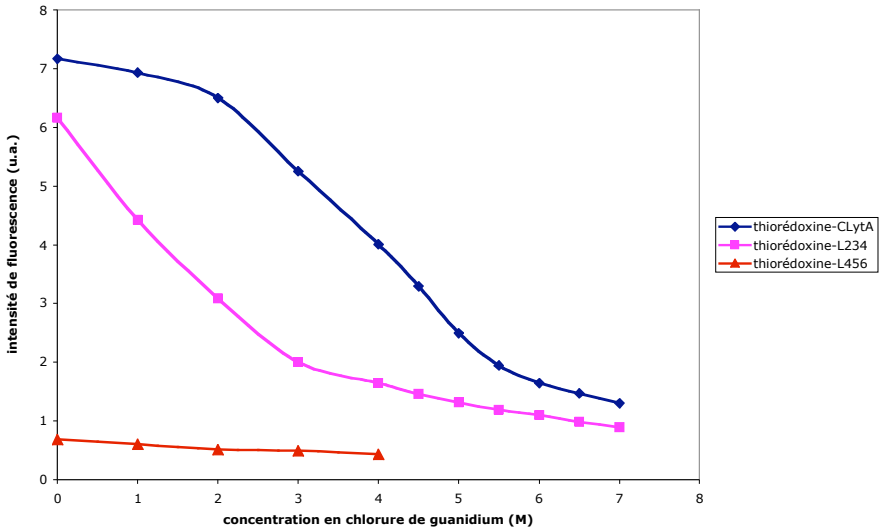


Figure 24 Mesure de la fluorescence des protéines de fusion thiorédoxine-ClytA, thiorédoxine-L456 et thiorédoxine-L234 en fonction de la concentration en chlorure de guanidium.
u.a. : unités arbitraires

Protéine	Nombres de *repeats*	Changement de fluorescence en présence de concentrations croissantes en chlorure de guanidium	concentration en chlorure de guanidium à partir de laquelle un changement d'intensité de fluorescence est est observé
thiorédoxine-CLytA	6	**oui**	**0 - 1 M**
thiorédoxine-L1234	4	**oui**	**0 - 1 M**
thiorédoxine-L2345	4	**oui**	**0 - 1 M**
thiorédoxine-L123	3	**oui**	**0 - 1 M**
thiorédoxine-L234	3	**oui**	**0 - 1 M**
thiorédoxine-L456	3	**non**	**0 - 1 M**
thiorédoxine-L23	2	**oui**	**0 - 1 M**
thiorédoxine	0	**oui**	**6 M**

Tableau 15 Mesures de la fluorescence des protéines de fusion thiorédoxine-fragment de domaine de liaison à 340 nm, en présence de concentrations croissantes en chlorure de guanidium.

Nous constatons que, excepté la thiorédoxine-L456, toutes les protéines de fusion testées présentent un changement de fluorescence en présence de concentrations croissantes en agent dénaturant. Il semble donc que la thiorédoxine-L456, présente dans la fraction soluble d'*E. coli* et pré-purifiée sur colonne de chélation, ne puisse conserver une structure. Pour les autres protéines de fusion, la diminution du signal de fluorescence, mesurée au cours de la dénaturation, atteste du passage des protéines étudiées d'un état natif à un état dénaturé.
Dans les conditions expérimentales que nous avons choisies (tampon PBS - pH 7,9), cette expérience permet aussi de mettre en évidence la faible stabilité des protéines de fusion au chlorure de guanidium puisque le changement d'intensité de fluorescence correspondant à un dépliement des protéines débute avec une concentration de l'ordre de 1M. Cette faible stabilité constitue un handicap pour les candidats *tag*s car ceux-ci ne permettront pas de purifier une protéine d'intérêt en conditions dénaturantes.

En résumé, la production de protéines de fusion thiorédoxine-domaine ClytA tronqué ainsi qu'une mesure de la fluorescence intrinsèque de ces protéines en présence d'un agent dénaturant nous ont permis de relever les observations suivantes :

- tous les fragments du domaine de liaison ClytA ne

semblent pas avoir la même capacité à adopter une structure lorsqu'ils sont fusionnés directement en C-terminal de la protéine reporter thiorédoxine. La thiorédoxine-L345 n'est pas produite dans la bactérie tandis que la thiorédoxine-L456 est produite en faible quantité soluble et ne présente pas de changement de fluorescence en présence de concentrations croissantes en agent dénaturant.

- quand elles sont resuspendues dans du tampon PBS pH 7,9, les protéines de fusion adoptant une structure se dénaturent en présence de faibles quantités de chlorure de guanidium, traduisant une faible stabilité face aux agents dénaturants, en l'absence de toute molécule stabilisante.

II.2. Analyse de l'affinité des fragments du domaine de liaison pour le DEAE-Sépharose Fast Flow

L'estimation d'affinité des protéines de fusion pour le DEAE-Sépharose Fast Flow a été réalisée en deux temps : d'abord sur les protéines de fusion préalablement purifiées sur colonne de chélation ; ensuite sur les protéines de fusion lorsque celles-ci sont présentes dans la fraction soluble d'*E. coli*.

II.2.1. Estimation de l'affinité pour le DEAE-Sépharose des protéines de fusion pré-purifiées

Une première estimation de l'affinité des protéines de fusion pour le DEAE-Sépharose a été réalisée de la façon suivante : après production dans *E. coli* et purification sur colonne de chélation (matériel et méthodes, paragraphes 4.2 et 5), les protéines thiorédoxine-domaine de liaison tronqué ClytA ont été déposées séparément sur colonne DEAE-Sépharose Fast Flow. Un premier lavage à l'aide d'un tampon phosphate 50 mM pH 7,9 a permis d'éliminer les éventuelles protéines ne s'accrochant pas sur la colonne. Puis, un gradient linéaire de NaCl 0-2 M a été appliqué de façon à éluer les protéines adsorbées sur la colonne par des interactions électrostatiques non spécifique. L'excès de NaCl présent sur la colonne a ensuite été éliminé par un second lavage au phosphate 50 mM puis un tampon phosphate 50 mM choline 2% a été appliqué. Seules les protéines adsorbées sur la colonne grâce à des interactions spécifiques entre le DEAE-Sépharose et les sites de liaison à la choline des fragments de domaine CLytA ont été éluées par compétition avec la choline. Ces dernières sont considérées comme possédant une affinité spécifique pour la choline et donc pour le DEAE-Sépharose.

L'affinité de toutes les protéines de fusion construites, excepté la thiorédoxine-L345 et la thiorédoxine-L456, a été estimée par ce protocole. Nous avons également testé l'affinité de la thiorédoxine seule pour le DEAE-Sépharose Fast Flow afin d'évaluer la contribution de la protéine reporter dans l'expérience. La thiorédoxine ne possède pas d'affinité spécifique pour le DEAE puisqu'elle s'élue en début de gradient NaCl (résultats non montrés).
L'estimation de l'affinité des protéines de fusion est reprise au tableau 16.

protéine	nombres de repeats	nombre de sites de liaison à la choline	affinité spécifique pour le DEAE-Sépharose Fast Flow
thio-CLytA	6	4 ou 5	+
thio-L1234	4	2 ou 3	+
thio-L2345	4	3	+
thio-L123	3	1 ou 2	+
thio-L234	3	2	+
thio-L23	2	1	-
thio	0	0	-

Tableau 16 Evaluation de l'affinité pour le DEAE des protéines de fusion thiorédoxine-fragment de domaine de liaison à la choline par chromatographie sur colonne DEAE-Sépharose Fast Flow.

Cette expérience de purification sur colonne DEAE-Sépharose permet de mettre en évidence les protéines de fusion thiorédoxine-L123 et thiorédoxine-L234. Ces deux protéines de fusion portent les plus petits fragments (trois *repeat*s) du domaine CLytA présentant une affinité spécifique pour le DEAE-Sépharose Fast Flow. Les figures 25 et 26 présentent d'une part le chromatogramme de l'expérience de purification sur colonne DEAE-Sépharose et d'autre par, l'analyse des fractions protéiques par Western Blot en utilisant soit un anticorps anti-*tag* 6 Histines soit un anticorps anti-ClytA (matériel et méthodes, paragraphe 10).

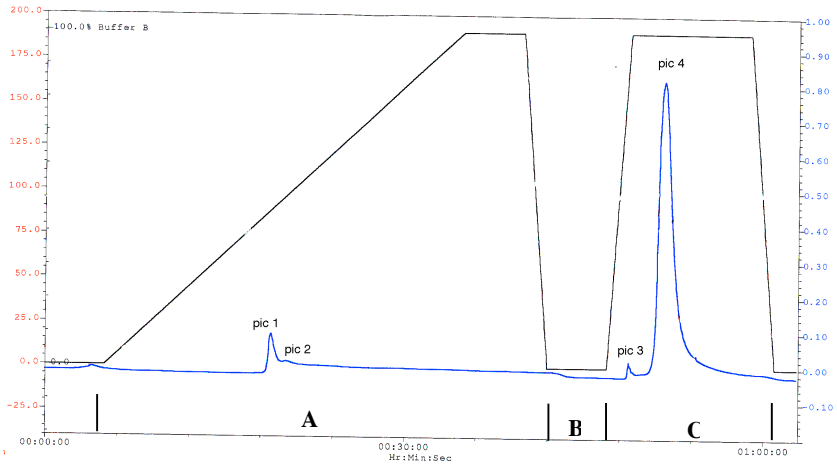


Figure 25 Profil chromatographique de la thiorédoxine-L234 sur colonne DEAE-Sépharose Fast Flow. La protéine de fusion est pré-purifiée sur colonne de chélation.
Zone A : gradient linéaire NaCl 0 - 2 M
Zone B : lavage phosphate 50 mM pH 7,9
Zone C : lavage phosphate 50 mM pH 7,9 - choline 2%
La ligne bleue correspond à la mesure de l'absorbance à 280 nm, mesurée à la sortie de la colonne.

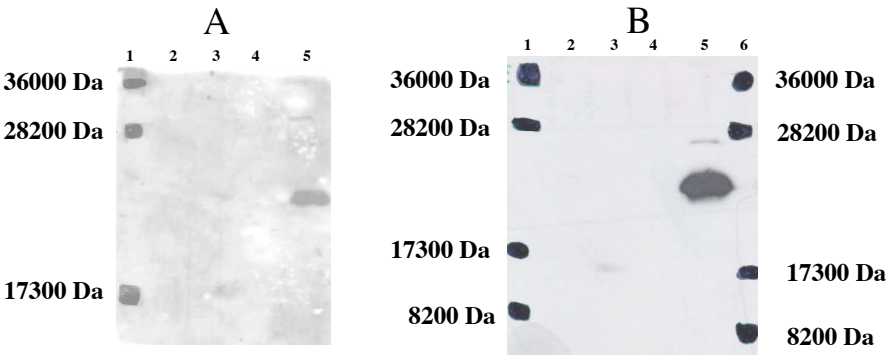


Figure 26 Analyse de l'affinité de la protéine thiorédoxine-L234 pour le DEAE-Sépharose par réaction de Western Blot sur les fractions protéiques éluées de la colonne DEAE-Sépharose Fast Flow.
A. : révélation de la thiorédoxine-L234 par un anticorps primaire anti-*tag* 6 Histidines
Piste 1 : marqueur de poids moléculaire
Pistes 2 et 3 : petits pics protéiques, numérotés 1 et 2, observés lors du gradient NaCl 0-2M
piste 4 : petit pic protéique observés lors du lavage phosphate 50 mM
Piste 5 : pic protéique observé lors du lavage au phosphate 50 mM choline 2%.

B : révélation de la thiorédoxine-L234 par un anticorps primaire anti-CLytA
Pistes 1 et 6 : marqueur de poids moléculaire
Pistes 2 à 4 : petits pics protéiques observés lors du gradient NaCl 0-2M et du lavage phosphate 50 mM
Piste 5 : pic protéique observé lors du lavage au phosphate 50 mM choline 2%.
Poids moléculaire attendu de la thiorédoxine-L234 : 22862 Da.

Pour la suite de notre travail, nous avons sélectionné le fragment L234 dont nous disposons des coordonnées tri-dimentionnelles complètes plutôt que le fragment L123 dont la structure du *repeat* 1 n'a jamais été résolue. Comme nous le verrons dans la quatrième partie des résultats, la sélection d'un fragment de structure connue permettra de définir des mutations stabilisantes par une approche bioinformatique, au départ de ses coordonnées tri-dimentionnelles.

II.2.2. Estimation de l'affinité pour le DEAE-Sépharose Fast Flow de la thiorédoxine-L234 présente dans la fraction soluble d'*E. coli*

Après avoir estimé l'affinité pour le DEAE-Sépharose de la thiorédoxine-L234 prépurifiée, nous avons évalué l'affinité de cette protéine lorsqu'elle est en présence d'autres polypeptides. Après surexpression de la protéine de fusion, la fraction soluble d'*E. coli* a donc été déposée sur colonne DEAE-

Sépharose Fast Flow et le même protocole de chromatographie que celui expliqué au paragraphe précédent a été appliqué.
En suivant l'absorbance à 280 nm à la sortie de la colonne, nous avons récolté les protéines s'éluant lors du lavage phosphate après injection de l'échantillon (*flow through*), lors de l'application du gradient NaCl 0-2M et lors du lavage à la choline 2%. Ces différentes fractions ont été analysées par SDS-PAGE après coloration au bleu de Coomasie et en Western Blot en utilisant l'anticorps anti-*tag* 6 Histidines (figure 27).
Nous constatons qu'une partie de la protéine de fusion se décroche dans le gradient NaCl avec les autres protéines solubles d'*E. coli* tandis qu'une autre partie de la protéine ne s'élue que par compétition avec la choline. Cette expérience a été réalisée en triplicat.
Afin d'avoir une meilleure estimation de la proportion de thiorédoxine-L234 présente dans les différentes fractions, nous avons réalisé un dosage de cette protéine par densitométrie après coloration au bleu de Coomasie (matériel et méthodes, paragraphe 9) . Dans cette expérience, 16% de la thiorédoxine-L234 s'élue lors du *flow through*, 39% lors du gradient NaCl 0-2M et 45% lors du lavage choline 2%. Ce dosage, réalisé une seule fois, devrait être répété afin d'obtenir une estimation plus précise.

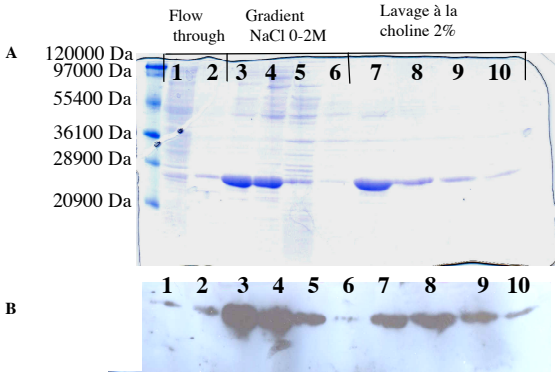


Figure 27 Analyse en SDS-PAGE et Western Blot des fractions protéiques s'éluant de la colonne DEAE-Sépharose Fast flow après application de différents tampons.
Pour chaque fraction de 5 ml récoltée, 10 µl ont été déposés par piste.
A : analyse en SDS-PAGE après coloration au bleu de Coomasie
Pistes 1 et 2 : fractions protéiques recueillies lors du premier lavage phosphate 50 mM après dépôt de l'échantillon (*flow through*)
Pistes 3 à 6 : fractions protéiques recueillies lors du gradient NaCl 0-2M
Pistes 7 à 10 : fractions protéiques collectées lors du lavage à la choline 2%

B : analyse en Western Blot en utilisant un anticorps anti-*tag* 6 Histidines
Pistes 1 et 2 : fractions protéiques recueillies lors du premier lavage phosphate 50 mM après dépôt de l'échantillon (*flow through*)
Pistes 3 à 6 : fractions protéiques recueillies lors du gradient NaCl 0-2M
Pistes 7 à 10 : fractions protéiques collectées lors du lavage à la choline 2%
Poids moléculaire attendu de la thiorédoxine-L234 : 22862 Da.

Il semble donc y avoir deux populations distinctes de la thiorédoxine-L234. Une première hypothèse permettant d'expliquer cette double population consiste à penser que la fusion de la thiorédoxine au candidat *tag* L234 ne permet pas à la protéine de fusion d'adopter une structure stable unique. L'adoption de plusieurs structures proches ou d'une structure peu stable pourrait conduire à une formation imparfaite des sites de liaison. Il s'en suivrait une diminution de l'affinité pour le DEAE, le non respect de la géométrie des sites de liaison ne permettant pas l'établissement de liaisons cation-π.
Outre le fait que la thiorédoxine-L234 est en compétition pour le DEAE-Sépharose avec un grand nombre de protéines solubles, une autre hypothèse permettant d'expliquer, en partie, la perte apparente d'affinité d'une partie de la population de thiorédoxine-L234 pourrait être la dimérisation du *tag*. Celle-ci rendrait les sites de liaison inaccessibles. Ces hypothèses seront réexaminées dans la discussion.

En résumé, la construction de protéines de fusion et l'estimation de l'affinité de ces protéines pour le DEAE-Sépharose Fast Flow nous a permis de sélectionner un fragment de domaine composé des *repeat*s 2, 3 et 4. Ce fragment présente une affinité spécifique pour le DEAE lorsque la protéine de fusion est prépurifiée sur colonne de chélation. Lorsqu'il est présent dans la fraction soluble d'*E. coli*, le candidat *tag* permet de purifier spécifiquement 40-50% de la thiorédoxine-L234.

II.3. Validation du fragment L234 en tant que *tag* de purification

II.3.1. Construction de la protéine de fusion MiaA-L234

Afin de vérifier que le fragment L234 pouvait être fusionné à une autre protéine tout en conservant sa capacité de liaison spécifique au DEAE-Sépharose, nous avons construit une deuxième protéine de fusion. Le gène codant pour la protéine MiaA a été amplifié par PCR au départ d'un plasmide aimablement fourni par l'Institut de Recherche Wiame, (CERIA, Bruxelles). L'amplicon a ensuite été cloné en lieu et place de la thiorédoxine aux sites *Asp*718 et *Sac*I du plasmide pet15b2-thiorédoxine-L234. La protéine de fusion obtenue se compose donc d'un *tag* 6 Histidines, de la protéine reporter MiaA, séparée par les résidus glycine et thréonine du fragment L234, comme c'était le cas pour la première protéine de fusion.

II.3.2. Estimation de l'affinité de la protéine de fusion MiaA-L234 pour le DEAE-Sépharose Fast Flow

La surexpression des protéines MiaA et MiaA-L234 a été réalisée selon le même protocole que celui utilisé pour la thiorédoxine-L234. Après avoir vérifié que la protéine MiaA seule ne possédait pas d'affinité spécifique pour le DEAE-Sépharose (résultats non illustrés), nous avons évalué l'affinité de MiaA-L234 pour cette matrice lorsque la protéine est présente dans la fraction soluble d'*E. coli.* En utilisant le même protocole de chromatographie que celui utilisé pour la protéine thiorédoxine-L234, nous avons constaté qu'une partie de la protéine s'éluait lors du gradient linéaire NaCl 0-2M alors qu'une autre partie de la protéine s'éluait uniquement en présence de choline, comme c'était le cas pour la thiorédoxine-L234 (figure 28). Cette expérience a été réalisée en triplicat.

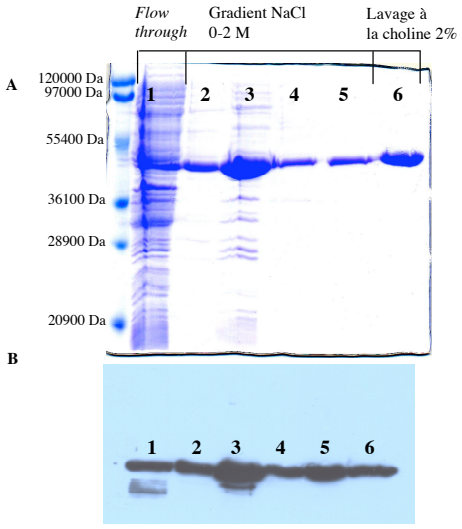


Figure 28 Estimation de l'affinité de MiaA-L234 pour le DEAE-Sépharose Fast Flow.
L'estimation a été réalisée par analyse en SDS-PAGE et Western Blot des fractions protéiques s'éluant de la colonne DEAE-Sépharose Fast flow après application de différents tampons.
A :analyse en SDS-PAGE après coloration au bleu de Coomasie
Piste 1 : fraction protéique recueillie lors du premier lavage phosphate 50 mM après dépôt de l'échantillon (*flow through*)
Pistes 2 à 5 : fractions protéiques recueillies lors du gradient NaCl 0-2M
Piste 6 : fraction protéique collectée lors du lavage à la choline 2%

B : analyse en Western Blot en utilisant un anticorps anti-*tag* 6 Histidines
Pistes 1 : fraction protéique recueillie lors du premier lavage phosphate 50 mM après dépôt de l'échantillon (*flow through*)
Pistes 2 à 5 : fractions protéiques recueillies lors du gradient NaCl 0-2M
Piste 6 : fraction protéique collectée lors du lavage à la choline 2%
Poids moléculaire attendu de MiaA-L234 : 46502 Da.

En conclusion, les expériences de purification de la protéine MiaA-L234 sur DEAE-Sépharose montrent que le fragment L234 peut également servir de *tag* de purification lorsqu'il est mis en fusion avec une autre protéine reporter que la thiorédoxine.

Au terme de cette deuxième partie des résultats, nous avons sélectionné un fragment du domaine de liaison à la choline ClytA, composé des motifs répétés 2, 3 et 4. Lorsque la thiorédoxine-L234 est pré-purifiée sur colonne de chélation puis déposée sur colonne DEAE-Sépharose Fast Flow, elle présente une forte affinité pour le DEAE puisqu'elle ne s'élue que par compétition avec la choline. Par contre, lorsqu'elle est présente dans la fraction soluble d'*E. coli*, le fragment L234 ne permet la purification que d'une partie de la protéine de fusion. Nous obtenons des résultats semblables lorsque le fragment L234 est fusionné en aval d'une autre protéine reporter, la protéine MiaA. Nous avons donc sélectionné un candidat prometteur mais, comme nous allons le voir aux points suivants, dont il faudrait améliorer certaines propriétés physico-chimiques.

PARTIE III : OPTIMALISATION DES CONDITIONS DE PRODUCTION DE LA PROTEINE THIOREDOXINE-L234

Avant de poursuivre notre étude sur le candidat *tag*, nous avons voulu optimiser les conditions de production de la thiorédoxine-L234. En effet, la protéine de fusion précipite majoritairement sous forme de corps d'inclusion lorsqu'elle est surexprimée dans des bactéries BL21(λDE3) pendant 4 heures à 37°C (conditions standards de surexpression, matériel et méthode, paragraphe 4.2). La figure 29 illustre la différence de solubilité entre la thiorédoxine et la thiorédoxine-L234. Une quantification par densitométrie des bandes protéiques colorée au bleu de Coomasie permet d'estimer que 68% de la thiorédoxine-L234 retrouve sous forme insoluble. Des résultats semblables sont observés pour la protéine MiaA-L234 (résultats non montrés).

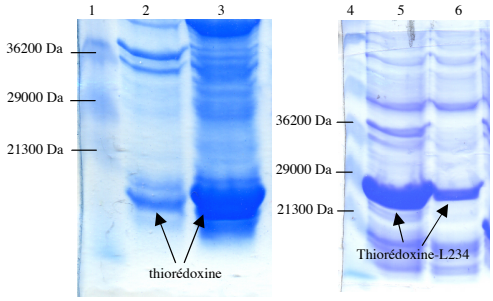


Figure 29 Mise en évidence des proportions relatives de thiorédoxine et de thiorédoxine-L234 dans les fractions insoluble et soluble *d'Escherichia coli* . Les productions de protéines ont été réalisées selon le protocole standard préconisé par la firme Novagen (induction à l'IPTG 1 mM pendant 4 h à 37°C). Les bandes protéiques sont mises en évidence par coloration au Bleu de Coomasie après migration sur gel SDS-PAGE 12%.
Piste 1 : marqueur de poids moléculaire
Piste 2 : fraction insoluble d'*E. coli* correspondant à l'équivalent de 200 µl de culture, après surexpression de la thiorédoxine pendant 4 h à 37°C.
Piste 3 : fraction soluble d'*E. coli* correspondant à l'équivalent de 200 µl de culture, après surexpression de la thiorédoxine pendant 4 h à 37°C.
Piste 4 : fraction insoluble d'*E. coli* correspondant à l'équivalent de 200 µl de culture, après surexpression de la thiorédoxine-L234 pendant 4 h à 37°C.
Piste 5 : fraction soluble d'*E. coli* correspondant à l'équivalent de 200 µl de culture, après surexpression de la thiorédoxine -L234 pendant 4 h à 37°C.
Piste 6 : marqueur de poids moléculaire
Poids moléculaire attendu de la thiorédoxine : 14857 Da
Poids moléculaire attendu de la thiorédoxine-L234 : 22862 Da

Ces données suggèrent que la fusion de la protéine reporter au candidat *tag* est responsable de l'insolubilisation de la thioédoxine-L234. Dans le contexte de la création d'un *tag* de purification à vocation industrielle, il est important de remédier à ce problème, la renaturation d'une protéine précipitée n'étant jamais assurée. L'augmentation de la quantité de protéines solubles peut se faire par deux approches : mise au point de conditions de production de la protéine d'intérêt et analyse de sa séquence en vue de définir des mutations ponctuelles permettant d'améliorer les caractéristiques physico-chimiques de la protéine.

Dans cette partie du travail, nous avons tenté d'optimaliser les conditions de production. La recherche de mutations jouant potentiellement un rôle sur la solubilité/stabilité de la thiorédoxine-L234 sera discutée dans la quatrième partie des résultats.

La précipitation d'une protéine d'intérêt sous forme de corps d'inclusion peut avoir plusieurs origines :

- une concentration protéique trop élevée dans la cellule, avec pour conséquence l'atteinte de la concentration limite de solubilité de la protéine d'intérêt,
- l'association d'intermédiaires partiellement foldés, conduisant à l'agrégation protéique.

L'importance des résidus hydrophobes lors du repliement d'une protéine est souligné dans la littérature. En effet, Georgiou et ses collaborateurs ont montré l'existence de barrières cinétiques lors du repliement d'une protéine (Georgiou and Valax, 1996). Ces barrières conduisent à l'accumulation d'espèces partiellement foldées ou d'intermédiaires foldés, présentant beaucoup de structures secondaires. Dans des expériences menées *in vitro*, ces auteurs ont montré que la présence de zones hydrophobes en surface des intermédiaires favorise leur auto-association qui conduit,à son tour, à l'agrégation protéique et donc à la formation de corps d'inclusion. Il est possible de jouer sur ce mécanisme sans changer la séquence de la protéine. En effet, dans la cellule, le mauvais repliement d'une protéine est minimisé notamment par l'activité des chaperones. Ces protéines ont la capacité de se lier à des intermédiaires partiellement foldés et les aident à atteindre leur structure tertiaire native, empêchant ainsi leur agrégation (Schlieker et al., 2002). En co-exprimant des chaperones ou en induisant leur production accrue suite à l'application d'un stress, on peut donc améliorer la proportion de protéines d'intérêt correctement repliées (Georgiou and Valax, 1996). Les stress sur la cellule peuvent être appliqués en ajoutant dans le milieu de culture certains consitutants tels que de l'éthanol, des sels en concentration élevée, sucre non métabolisable ou en provoquant un choc de température.

La diminution de la concentration en protéines peut se réaliser notamment en faisant varier la température de culture lors de l'induction (modification de la vitesse de synthèse protéique) et/ou en diminuant la concentration en inducteur (diminution de la fréquence de synthèse protéique).

Dans un premier temps, nous avons d'abord fait varier la température de culture. Trois températures ont été testées : 37°C, 24°C et 18°C (matériel et méthode, paragraphe 4.3). Pour chaque température, la production de la thiorédoxine-L234 a été induite par ajout d'IPTG 0,5 mM, au départ d'une préculture incubée à 37°C. Après 1h, 2h, 3h, 4h, 7h, 9h et 21h d'induction, 20 ml de culture ont été prélevés pour chaque température testée. Les fractions soluble et insoluble de chaque échantillon ont été analysées par Western Blot en utilisant comme anticorps primaire un anticorps monoclonal anti-*tag* 6 Histidines (3G12) et comme anticorps secondaire un anticorps anti-souris couplé à la peroxydase. La révélation des protéines reconnues par l'anticorps primaire se fait par chémoluminescence.
La figure 30 reprend les résultats obtenus. Nous observons une forte influence de la température d'induction sur la solubilité de la thiorédoxine-L234. Les meilleures conditions sont soit 7h d'induction à 24°C soit 21h d'induction à 18°C.

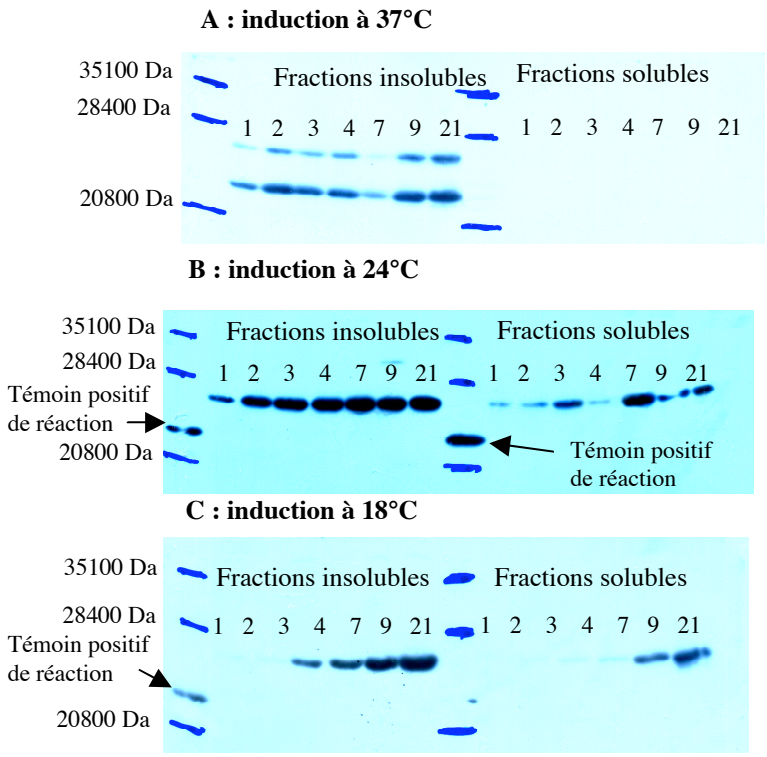


Figure 30 Influence de la température de culture sur la solubilité de la thiorédoxine-L234
Après 1, 2, 3, 4, 7, 9 et 21 h d'induction à l'IPTG 1 mM, 20 ml de culture sont prélevés et centrifugés. Les fractions solubles et insolubles sont préparées et déposées sur gel SDS-PAGE 12%. La thiorédoxine-L234 est mise en évidence par réaction de Western Blot en utilisant comme anticorps primaire un anticorps anti-*tag* 6 Histidines et comme anticorps secondaire, un anticorps anti-souris couplé à la peroxydase.

Nous avons ensuite testé l'influence de la concentration en inducteur sur la solubilité de la thiorédoxine-L234 à 18°C (matériel et méthodes, paragraphe 4.4). Les inductions ont été réalisées par ajout d'IPTG 1mM, 0,5 mM, 0,1 mM et 0,05 mM. 20 ml de culture ont été prélevés après 7h, 9h et 21h d'induction pour chaque concentration en inducteur testée. Les fractions soluble et insoluble ont ensuite été analysées par Western Blot. Les résultats obtenus révèlent que, lors d'une induction à 18°C, il est préférable de garder une concentration élevée (1 mM) en inducteur (résultats non illustrés). Le reste de chaque culture a ensuite été centrifugé, la protéine thiorédoxine-L234 purifiée sur colonne de chélation à partir de la fraction soluble et dosée par la méthode BCA (Smith et al., 1985) (matériel et méthodes, paragraphe

6). Le dosage de la thiorédoxine-L234 soluble pour chaque concentration en IPTG donne la même tendance (tableau 17).

	Quantité de thiorédoxine-L234 soluble dans 240 ml de culture (mg)	Extrapolation de la quantité de thiorédoxine-L234 soluble produite par litre de culture (mg/l)
Induction à l'IPTG 1 mM	14,2	59,2
Induction à l'IPTG 0,5 mM	8,9	37,1
Induction à l'IPTG 0,1 mM	4	16,7
Induction à l'IPTG 0,05 mM	0,6	2,5

Tableau 17 Quantité de thiorédoxine-L234 soluble produite après 21h d'induction à 18°C au départ de cultures de 300 ml et en faisant varier la concentration en IPTG.

Ayant défini deux paramètres modulant la concentration en protéines produites, nous avons ensuite essayé de moduler la production de protéines chaperones. Ces dernières peuvent être induites en modifiant la composition du milieu de culture. En effet, d'après la littérature, l'application d'un choc osmotique (NaCl ou sucrose) ainsi que l'ajout d'éthanol dans le milieu de culture peut induire la production de protéines chaperones natives (Brissette et al., 1990) (Ma et al., 1996) (Bowden and Georgiou, 1990). De même, le remplacement du milieu LB classique par un milieu plus riche a déjà permis de produire de plus grandes quantités de protéines solubles lors d'essais de surexpression (Song et al., 1999) (Lee et al., 1998).
En pratique, nous avons ensemencé six cultures de 300 ml contenant les milieux suivants :
- LB (contrôle négatif : 1% bactotryptone, 0,5% Yeast extract, 1% NaCl),
- LB + NaCl 0,3M, soit NaCl 0,45M final (Bowden and Georgiou, 1990),
- LB + sucrose 0,6M (Bowden and Georgiou, 1990),
- LB + éthanol 4% (Brissette et al., 1990),
- Milieu très riche (2% bactopeptone, 0,2% Na_2HPO_4, 0,1% KH_2PO_4, 0,8% NaCl, 1,5% yeast extract, 0,2% glucose) (Lee et al., 1998).

20 ml de culture ont été prélevés après 7h, 9h et 21h d'induction à 18°C, en utilisant une concentration en IPTG de 1 mM. L'analyse des fractions soluble et insoluble par Western Blot montre que la quantité de thiorédoxine-L234 soluble est plus importante lorsque l'induction se fait dans du LB +NaCl 0,3M (figure 31).

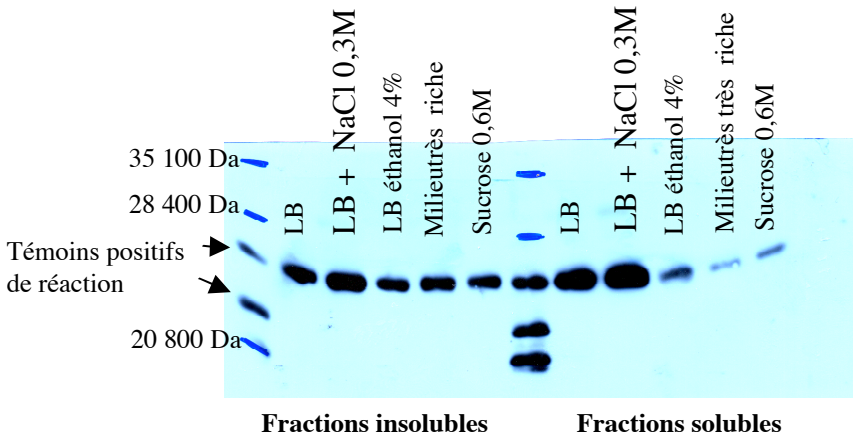


Figure 31 Analyse par Western Blot des fractions soluble et insoluble de la thiorédoxine-L234 après production 21 h à 18°C dans différents milieux. Les fractions solubles et insolubles sont préparées et déposées sur gel SDS-PAGE 12%. La thiorédoxine-L234 est mise en évidence par réaction de Western-Blot en utilisant comme anticorps primaire un anticorps anti-tag 6 Histidines et comme anticorps secondaire, un anticorps anti-souris couplé à la peroxydase.

Les quantités protéiques calculées sur des volumes de culture de 300 ml montrent qu'il y a une augmentation de la production de la thiorédoxine-L234 de l'ordre de 18% lorsqu'on cultive les bactéries BL21(λDE3) dans du milieu LB + NaCl 0,3M au lieu du milieu LB (tableau 18).

volume de culture	milieu de culture	extrapolation de la quantité de protéines solubles (mg/l)	moyenne de la quantité obtenue (mg/l)
300 ml	LB	52	
300 ml	LB	51	**53 +/- 3**
300 ml	LB	58	
300 ml	LB + NaCl 0,3M	74,5	
300 ml	LB + NaCl 0,3M	73,5	**72 +/- 2**
300 ml	LB + NaCl 0,3M	69,9	

Tableau 18 Quantité protéique obtenue pour la thiorédoxine-L234 lorsque l'on fait varier la composition du milieu de culture.

Le changement de composition du milieu de culture par ajout de NaCl 0,3M a donc une influence sur la quantité de thiorédoxine-L234 soluble produite.

En résumé, au terme de cette deuxième partie expérimentale, nous avons mis au point des conditions de surexpression permettant d'optimiser la production de thiorédoxine-L234. Nous obtenons une plus grande quantité de la protéine d'intérêt en cultivant les bactéries BL21(λDE3) dans du milieu LB additionné de NaCl 0,3M et en réalisant la surexpression par ajout d'IPTG 1 mM pendant 21h à 18°C.

PARTIE IV : DEFINITION DE MUTATIONS PONCTUELLES EN VUE D'AMELIORER LA SOLUBILITE / STABILITE DU *TAG* ET ANALYSE DES MUTANTS

Les deux premières étapes expérimentales dans l'élaboration d'un *tag* de purification ont consisté à sélectionner le plus petit fragment possible du domaine de liaison ClytA présentant une affinité spécifique pour le DEAE et à définir des conditions de production permettant d'améliorer la quantité de thiorédoxine-L234 soluble produite.
Mais, comme nous l'avons déjà souligné, un désavantage du candidat *tag* semble être sa propension à faire précipiter la protéine à laquelle il est fusionné. Le candidat L234 présente donc un défaut de solubilité. De plus, lorsque la protéine de fusion est présente dans la fraction soluble d'*E. coli*, elle semble se diviser en deux populations se distinguant l'une de l'autre par leur affinité pour le DEAE-Sépharose. Ce comportement pourrait s'expliquer, entre autres, par l'adoption d'une structure peu stable avec formation de sites de liaison à la choline suboptimaux.
Partant de ces observations, nous avons choisi trois approches bioinformatiques permettant de définir des mutations ponctuelles, pouvant avoir un effet sur la solubilité ou la stabilité du fragment L234 lorsqu'il est mis en fusion avec la thiorédoxine.
Ces trois approches sont :

- recherche de résidus dont l'accessibilité au solvant change lors du raccourcissement du domaine de liaison ClytA,
- comparaison de la séquence des *repeat*s 2, 3 et 4 au consensus de séquence d'un *repeat*, défini sur base de l'alignement de 126 *repeat*s de domaines de liaison à la choline,
- analyse de la séquence des *repeat*s 2, 3 et 4 par l'algorithme PoPMuSiC (Gilis and Rooman, 2000).

IV.1. Définition de mutations ponctuelles

IV.1.1. Recherche de résidus hydrophobes exposés au solvant suite au raccourcissement du domaine de liaison ClytA

Selon Georgiou, lors du repliement d'une protéine, la présence de zones hydrophobes en surface d'intermédiaires partiellement foldés favorise leur auto-association (Georgiou and Valax, 1996). Ce phénomène conduit à l'agrégation protéique *in vitro* et donc à la formation de corps d'inclusion.
Dans le cadre de notre recherche, le fait de tronquer le domaine de liaison ClytA, contenant 6 *repeat*s, pour le réduire aux *repeat*s 2, 3 et 4, conduit probablement à l'exposition au solvant de résidus hydrophobes. Il est donc

intéressant de rechercher ces résidus et de les muter, en espérant ainsi obtenir une augmentation de la solubilité globale.

La mutation de résidus hydrophobes en résidus hydrophiles pourrait également avoir une influence sur la stabilité de la thiorédoxine-L234, comme cela a été décrit dans la littérature (Strub et al., 2004) (Banci et al., 2003). L'origine de l'augmentation de stabilité suite au remplacement d'un résidu hydrophobe par un résidu hydrophile peut trouver son origine dans la création de nouveaux ponts hydrogènes entre le solvant et le résidu nouvellement introduit, conduisant ainsi à une amélioration de la stabilité (Pedone et al., 2001). Elle peut également trouver son origine dans un changement de l'entropie de la protéine suite à l'introduction de la mutation, affectant de ce fait sa stabilité.

En résumé, que l'on se place du point de vue de la solubilité ou de la stabilité de la protéine de fusion, il semble intéressant de mettre en évidence des résidus hydrophobes potentiellement exposés au solvant et de les muter.

Nous avons donc soumis les coordonnées tri-dimentionnelles du domaine de liaison complet et des *repeat*s 2 ,3 et 4 à deux programmes calculant, entre autres, l'accessibilité au solvant de chaque résidu : DSSP (Kabsch and Sander, 1983) et ENVIRON (Koehl and Delarue, 1994). Nous avons ensuite comparé l'accessibilité au solvant entre la structure complète et la structure tronquée pour chaque résidu (figure 32).

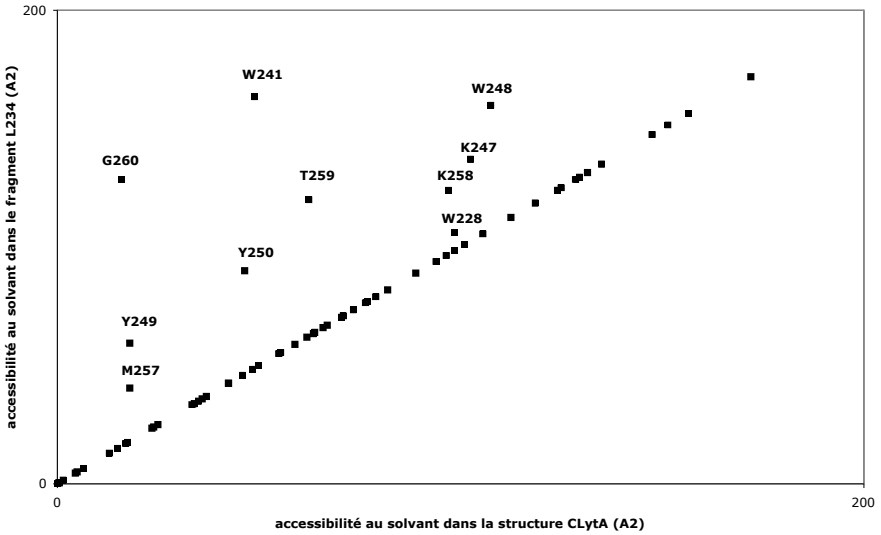


Figure 32 Mise en évidence de résidus hydrophobes exposés au solvant suite à un raccourcissement du domaine de liaison ClytA.

Nous observons une augmentation de l'accessibilité au solvant pour les résidus hydrophobes W241, W248, Y 249, Y250 et M257.

Les résidus Y 249 et M 257 participent à la formation d'un des deux sites de liaison à la choline présents sur les *repeat*s 2, 3 et 4. Ils ne sont donc pas

mutables. Par contre, les résidus W241, W248 et Y250 ne participent pas à la formation des cages d'aromatiques dans cette portion du domaine. Nous avons décidé de les muter respectivement en acide glutamique (E), acide aspartique (D) et acide glutamique (E). Le choix de ces résidus a pour objectif de remplacer les résidus aromatiques par des acides aminés possédant une chaîne latérale assez volumineuse pour éviter la création de vides stériques. Le fait qu'ils soient chargés négativement ne nous semble à priori pas incompatible avec l'amélioration du *tag*. En effet, le support chromatographique utilisé étant chargé positivement au pH auquel nous travaillons, les charges négatives introduites pourraient renforcer l'affinité globale du *tag* pour le DEAE-Sépharose. Il faut cependant noter que l'introduction de ces résidus a pour conséquence logique un changement de point isoélectrique de la thiorédoxine-L234. Le choix de ces résidus chargés sera rediscuté au paragraphe IV.2.2 et dans la discussion générale.

IV.1.2. Sélection de mutations par comparaison de la séquence des repeats 2, 3 et 4 au consensus de séquence d'un repeat

Dans une seconde approche, nous avons comparé la séquence des *repeat*s 2, 3 et 4 au consensus, position par position, pour mettre en évidence des classes d'acides aminés qui sont statistiquement sur- ou sous-représentées par rapport à celles définies dans le consensus (partie I des résultats). En parallèle, nous avons comparé la préférence de structure secondaire du résidu présent dans la séquence à la structure secondaire réellement adoptée à cette position dans la structure.
Le tableau 19 présente une partie des résultats obtenus. Nous pouvons, par exemple, observer qu'à la position 4 du *repeat* 2, un glutamate est présent alors que, selon les programmes d'alignement, nous devrions trouver à cette position un résidu hydrophobe dans 61,9% à 72,3% des cas. De plus, la structure secondaire préférentiellement adoptée par le glutamate est une hélice α alors que, dans la structure du domaine de liaison ClytA, ce résidu participe à la formation d'un brin β. Dans ce cas-ci, ni la classe de résidus, ni la préférence de structure secondaire n'étant en concordance avec le consensus et la structure réelle, ce candidat est intéressant pour la mutagenèse.

position dans le repeat	n° du résidu	résultats alignements de 126 repeats par 4 programmes	structure secondaire réelle	préférence de structure secondaire du résidu de la séquence			résidus proposé	préférence de structure secondaire du résidu proposé		
				hélice α	brin β	turn		hélice α	brin β	turn
1	D197	75,2 à 80,9 % polaires dont 62,4 à 63,5 T	loop	0,99	0,39	1,24	T	0,76	1,17	0,9
2	K198	63 à 66,7 % de glycines	brin	1,23	0,69	1,07	G	0,43	0,58	1,77
3	F199	92,9 à 100 % aromatiques dont 81 à 92,8% de W	brin	1,16	1,33	0,59	pas de mutation			
4	E200	61,9 à 72,3% hydrophobes dont 27,8 à 29,4 % de L et 22,2 à 27,8% de V	brin	1,59	0,52	1,01	L V	1,34 0,9	1,22 1,87	0,57 0,41
5	K201	42,1 à 44,8 % de chargés dont 35,7 à 38,4 % de K 44,4 à 48,4 % de polaires dont 30,4 à 34,1% de Q	brin	1,23	0,69	1,07	pas de mutation			
6	I202	39,2 à 44,4% de chargés dont 29,4 à 34,9% de D 22,4 à 24 % d'hydrophobes dont V > I > L > A 17,5 à 20,8 % de polaires dont 11,9 à 12,8% de N 15,1 à 16 % de Y	brin	1,09	1,67	0,47	pas de mutation			
7	N203	47 à 48,4 % de polaires dont 38,4 à 38,9 % de N 24,6 à 30% de chargés dont 16,7 à 19% de K	turn	0,76	0,48	1,34	pas de mutation			
8	G204	56 à 63,5% de glycines 18,3 à 24% de chargés dont 11,9 à 13,6% de D	turn	0,43	0,58	1,77	pas de mutation			

Tableau 19 Comparaison des huit premiers résidus du repeat 2 aux huit premières positions du consensus.
La structure préférenciellement adoptée par un résidu est une valeur de fréquence normalisée, calculée à partir de la proportion de chaque résidu que l'on retrouve dans une conformation particulière, divisée par cette proportion pour tous les résidus (Creighton, 1993) (Williams et al., 1987)1987) (Wilmot and Thornton, 1988)

En alignant les *repeat*s du fragment L234 par les programmes ClustalW (Thompson et al., 1994) et MatchBox (Depiereux and Feytmans, 1992), nous avons constaté qu'à certaines positions, la classe d'acides aminés présente dans le fragment de domaine ne correspondait pas à celle définie dans le consensus (tableau 20).

position dans le repeat	résidus présents dans les repeats 2, 3 et 4 du domaine de liaison CLytA	résidus présents dans le consensus	caractéristique des repeats resélectionnés pour réaliser un nouvel alignement.
position 2	deux résidus chargés (K) deux glycines	glycines	repeats présentant une lysine en position 6
position 4	trois résidus chargés (deux positifs et deux négatifs)	résidus hydrophobes	repeats présentant un résidu chargé en position 8
position 8	deux acides aspartiques et une glycine	glycines	repeats présentant un résidu chargé en position 12

Tableau 20 Liste des résidus exceptionnellement présents dans les *repeat*s 2, 3 et 4 du domaine de liaison ClytA par rapport au consensus de *repeat* établi.

Dans un tel cas de figure, il est possible que la présence d'une classe particulière de résidus s'explique par l'établissement d'interactions entre ces résidus et des acides aminés d'une autre (ou même) classe, également présents, à titre exceptionnel, à une autre position du consensus. Si situation pourrait signifier l'existence de mutations compensatoires dans la séquence du fragment L234. Afin de tester cette hypothèse, chaque position s'écartant du consensus a été analysée. Nous pouvons prendre comme exemple la position 8 à laquelle nous retrouvons des résidus chargés alors que l'on trouve normalement des résidus hydrophobes à cette position. Parmi les 126 *repeat*s utilisés pour définir le consensus de séquence (paragraphe 1.2.3), les *repeat*s présentant un résidu chargé ont été sélectionnés et alignés par les programmes ClustalW et MatchBox. Puis, les classes d'acides aminés présentes à chaque position ont été définies et comparées au consensus. Dans aucun cas, nous n'avons pu corréler la présence des classes particulières de résidus aux positions 2, 4 et 8 des *repeat*s L234 à la présence d'une autre classe particulière de résidus à une autre position dans les *repeats*.
Finalement, ayant écarté l'existence de mutations compensatoires, la comparaison de la séquence des *repeat*s 2, 3 et 4 au consensus de *repeat*s nous a permis de sélectionner 14 résidus potentiellement mutables (tableau 21).

position	1	2	3	4	5	6	7	7'	8	9	10	11	12	13	14	15	16	17	18	18'	19	20
séquence du repeat 1 proposition de mutation														V	H	S	D	G	S	Y	**P M**	**K A**
séquence du repeat 2 proposition de mutation	**D T**	**K G**	F	**E V**	K	I	N		G	T	W	Y	Y	F	D	S	S	G	**Y S, D**	M	L	A
séquence du repeat 3 proposition de mutation	**D T**	**R G**	W	**R V**	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E		M	A
séquence du repeat 4 proposition de mutation	T	G	W	**K V**	K	I	**A N, K**		**D G**	**K S**	W	Y	Y	F	N	E	E	G	A		M	**K A**
séquence du repeat 5 proposition de mutation	T	G																				
consensus général d'un repeat	pol/X	G/X	arom	phobe/X	ch/pol	X	ch/pol		G	X	W	Y	Y	phobe	ch/pol	petit AA	ch/pol	G	petit AA		phobe	phobe

Tableau 21 Représentation des 14 résidus de la séquence du *tag* L234 différant des classes de résidus définies dans le consensus de *repeat* soit par leurs propriétés physico-chimiques soit par la structure secondaire qu'ils adoptent préférentiellement.
Les lettres majuscules correspondent aux symboles des 20 acides aminés
Les positions 2', 4' et 11' sont des positions supplémentaires, non définies dans le consensus général de *repeat*s, car la séquence répétée analysée possède plus de 20 résidus.
Dans chaque encadré, le résidu supérieur correspond à l'acide aminé réellement présent dans la séquence tandis que le résidu inférieur correspond à la mutation proposée.

Vu le nombre de résidus ne répondant pas au consensus, il semble difficile de sélectionner certaines mutations plutôt que d'autres. Cette analyse reste cependant un point de comparaison que nous pourrons utiliser lors de l'analyse de la séquence du fragment de domaine L234 par l'algorithme PoPMuSiC (voir point suivant).

IV.1.3. Sélections de mutations stabilisantes par l'algorithme PoPMuSiC

L'analyse de la séquence des *repeat*s 2, 3 et 4 ainsi que la recherche de résidus hydrophobes exposés au solvant nous ont permis de mettre en évidence 17 résidus potentiellement mutables.
Toujours dans une optique de stabilisation du fragment L234, nous avons utilisé une troisième approche en soumettant la séquence des *repeat*s 2, 3 et 4 à l'algorithme PoPMuSiC (Gilis and Rooman, 2000).
De manière générale, PoPMuSiC évalue le changement de stabilité d'une protéine ou d'un peptide après introduction de toutes les mutations ponctuelles possibles, soit dans une séquence complète, soit dans une région spécifiée par l'utilisateur. Il fournit une liste des mutations les plus déstabilisantes, des mutations les plus stabilisantes ainsi qu'une liste de mutations neutres. Pour ce faire, il utilise des potentiels de distance et des potentiels de torsion dérivés d'une banque de structures, la combinaison des potentiels utilisés étant fonction de l'accessibilité au solvant du résidu considéré.

De façon plus détaillée, les potentiels de torsion décrivent uniquement des interactions locales le long de la séquence. Ils tiennent compte de la propensions des domaines (ϕ, Ψ et ω) ou de paires de domaines (ϕ, Ψ et ω) à être associé à un résidu donné. Les potentiels de distance sont dominés par des interactions non-locales. Ils sont basés sur la propension de paires de résidus (a_i, a_j), situés aux positions i et j de la séquence, à être séparés par une distance spaciale $d_{i,j}$, calculée entre les centroïdes moyens des chaînes latérales des résidus considérés.

L'évaluation du changement d'énergie libre de repliement après introduction d'une mutation est estimée en faisant la différence entre l'énergie libre de repliement dans la structure mutée et l'énergie libre de repliement dans la structure native.

Un changement d'énergie libre de repliement négatif implique que la mutation est stabilisante, tandis que s'il est positif, la mutation est déstabilisante.

L'utilisation de PoPMuSiC constitue donc une approche enrichissante supplémentaire puisqu'elle permet de tenir compte de critères structuraux, ce qui n'est pas le cas de l'analyse par comparaison à un consensus, celle-ci ne faisant intervenir que des fréquences de résidus à des positions particulières d'une séquence.

En pratique, les coordonnées tri-dimentionnelles des *repeat*s 2, 3 et 4 du domaine de liaison ClytA ont été soumises à l'algorithme PoPMuSiC et les résultats ont été analysés au laboratoire d'Ingénierie Biomoléculaire de l'Université Libre de Bruxelles (Dr Dimitri Gilis et Dr Marianne Rooman).

Les critères d'application et d'analyse du programme ont été définis comme suit :

- utilisation de deux banques de données différentes pour dériver les potentiels de torsion et de distance,
- l'amidase de *Streptococcus pneumoniae* (LytA) étant un homodimère, utilisation des coordonnées tri-dimentionnelles des *repeat*s 2, 3 et 4 des chaînes A et B du dimère,
- pas d'essai de mutation avec les résidus cystéine et proline qui peuvent introduire des modifications importantes dans la structure,
- cohérence des changements d'énergie libre calculés en dérivant les potentiels des deux banques de données,
- cohérence des changements d'énergie libre calculés pour les chaînes A et B.

Les résidus potentiellement stabilisants sélectionnés par l'agorithme PoPMuSiC sont schématisés au tableau 22 et comparés aux mutations proposées par l'analyse du consensus.

position	1	2	3	4	5	6	7	7'	8	9	10	11	12	13	14	15	16	17	18	18'	19	20
séquence du repeat 1 mutations proposées par l'algorithme PoPMuSiC														V	H	S	D	G	S	Y	P	K
mutations proposées après comparaison au consensus																					M	A
séquence du repeat 2 mutations proposées par l'algorithme PoPMuSiC	D	K	F	E L,V	K	I	N		G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
mutations proposées après comparaison au consensus	T	G		V															S, D			
séquence du repeat 3 mutations proposées par l'algorithme PoPMuSiC	D	R	W	R L,V	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E		M	A
mutations proposées après comparaison au consensus	T	G		V																		
séquence du repeat 4 mutations proposées par l'algorithme PoPMuSiC	T	G	W	K L, T, V	K	I	A N		D G	K	W	Y	Y	F	N	E	E	G	A		M	K
mutations proposées après comparaison au consensus				V			N, K		G	S												A
séquence du repeat 5 mutations proposées par l'algorithme PoPMuSiC	T	G																				
mutations proposées après comparaison au consensus																						
consensus général d'un repeat	pol/X	G/X	arom	phobe/X	ch/pol	X	ch/pol		G	X	W	Y	Y	phobe	ch/pol	petit AA	ch/pol	G	petit AA		phobe	phobe

Tableau 22 Représentation schématique des mutations définies par comparaison de la séquence des *repeat*s 2, 3 et 4 au consensus de *repeat* et des mutations stabilisantes définies par l'algorithme PoPMuSiC.
Les résidus surlignés en gris correspondent à des résidus proposés simultanément par l'algorithme PoPMuSiC et par la comparaison au consensus de séquences.

Les données obtenues montrent que, pour cinq positions, il y a concordance des résultats entre l'analyse par le consensus et l'algorithme PoPMuSiC.
Ces positions à muter sont donc :
- les résidus chargés en position 4 des *repeat*s 2, 3 et 4, à muter préférentiellement en valine,
- les résidus en position 7 et 8 du troisième *repeat*, à remplacer respectivement par une asparagine (N) et une glycine (G).

IV.1.4. Définition de trois mutants

Les résultats des trois types d'analyses bioinformatiques nous permettent de créer trois mutants.
Le premier concerne les trois résidus hydrophobes plus exposés au solvant suite au raccourcissement du domaine de liaison ClytA. Bien que ces mutants ne soient pas repris dans les résultats de l'algorithme PoPMuSiC, il nous a semblé intéressant de tester cette hypothèse. Ce mutant est appelé par la suite thiorédoxine-EDE-L234 (tableau 23).

position	1	2	3	4	5	6	7	7'	8	9	10	11	12	13	14	15	16	17	18	18'	19	20
REPEAT 1														V	H	S	D	G	S	Y	P	K
REPEAT 2	D	K	F	E	K	I	N		G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
REPEAT 3	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E		M	A
REPEAT 4	T	G	E	K	K	I	A		D	K	D	Y	E	F	N	E	E	G	A		M	K
REPEAT 5	T	G																				

Tableau 23 Représentation des mutations proposées pour les résidus hydrophobes exposés au solvant.

Le deuxième mutant (thiorédoxine-V1V2V3-L234) concerne les acides aminés chargés, situés dans les *repeat*s 2, 3 et 4 à la position 4 (tableau 24). Ces trois résidus ont été mutés en valine, seul résidu proposé en commun par l'analyse du consensus et par l'algorithme PoPMuSiC.

position	1	2	3	4	5	6	7	7'	8	9	10	11	12	13	14	15	16	17	18	18'	19	20
REPEAT 1														V	H	S	D	G	S	Y	P	K
REPEAT 2	D	K	F	V	K	I	N		G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
REPEAT 3	D	R	W	V	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E		M	A
REPEAT 4	T	G	W	V	K	I	A		D	K	W	Y	Y	F	N	E	E	G	A		M	K
REPEAT 5	T	G																				

Tableau 24 Représentation des mutations proposées pour les résidus chargés, situés en position 8 des *repeat*s 2, 3 et 4 du domaine de liaison ClytA.

Le troisième mutant concerne les résidus alanine et aspartate situés en positions 7 et 8 du repas 4. Ces acides aminés participent à la formation du *turn* situé entre deux brins β constituant une hairpin β. L'analyse basée sur la séquence du consensus ainsi que celle réalisée par l'algorithme PoPMuSiC proposent de muter ces deux résidus respectivement en asparagine et en glycine (tableau 25). Lorsqu'on analyse l'alignement de 126 *repeat*s par les programmes ClustalW (Thompson et al., 1994) et Match-Box (Depiereux and Feytmans, 1992), le couple NG apparaît plus fréquemment que les autres couples à ces positions dans les *repeat*s (résultats non montrés). Cette observation renforce l'hypothèse selon laquelle le couple de mutations NG est un bon candidat. Ce mutant a été appelé thiorédoxine-NG-L234.

position	1	2	3	4	5	6	7	7'	8	9	10	11	12	13	14	15	16	17	18	18'	19	20
REPEAT 1														V	H	S	D	G	S	Y	P	K
REPEAT 2	D	K	F	E	K	I	N		G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
REPEAT 3	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E		M	A
REPEAT 4	T	G	W	K	K	I	N		G	K	W	Y	Y	F	N	E	E	G	A		M	K
REPEAT 5	T	G																				

Tableau 25 Représentation des mutations proposées pour les résidus situés en positions 7 et 8 du *repeat* 4 du domaine de liaison ClytA.

Il faut souligner que l'algorithme PoPMuSiC a été développé de façon à proposer des mutations stabilisantes individuelles. Il ne garantit en aucun cas que deux mutations définies comme stabilisantes quand elles sont prises individuellement, ont un effet additif sur la stabilité globale de la protéine lorsqu'elles sont introduites ensemble. Nous avons quand même décidé de créer deux mutants portant plusieurs mutations définies par PoPMuSiC. Le regroupement des mutants valine en position 4 des trois *repeat*s nous a paru plausible en regard des résultats obtenus par l'analyse du consensus. En effet, lorsqu'on aligne 126 *repeat*s provenant de 19 domaines différents de liaison à la choline, 61,9 à 72,3% de résidus à cette position sont hydrophobes, dont 22,2% à 27,8% de valine. Vu l'absence apparente de mutations compensatoires dans la séquence permettant d'expliquer la présence de résidus chargés à ces positions, leur remplacement par des valines dans les trois *repeat*s est envisageable. De même, l'introduction des mutations N et G simultanément semble réaliste vu le nombre de fois que cette combinaison de résidus est observée à ces deux positions dans les 126 motifs répétés analysés.

IV.2. Construction et analyse des mutants thiorédoxine-EDE-L234, thiorédoxine-NG-L234 et thiorédoxine-V1V2V3-L234

IV.2.1. Construction des protéines de fusion mutantes

Les mutations ont été introduites dans la séquence des *repeat*s 2, 3 et 4 par réaction PCR en utilisant des oligonucléotides portant chacun une ou plusieurs mutations désirées. Les mutants ont donc été construits en plusieurs étapes. La stratégie générale de construction des mutants est reprise à la figure 33.
Dans un première étape, des oligonucléotides complémentaires des régions amont et aval de la protéine de fusion et des oligonucléotides portant les mutations désirées ont servi a amplifier en deux fragments le gène codant

pour la protéine de fusion. Puis, chaque région amplifiée possédant une portion de séquence complémentaire à l'autre région, les deux fragments d'ADN s'apparient naturellement à une température particulière. Cette zone d'appariement sert d'amorce à la Taq Polymérase pour continuer la synthèse de brins complémentaires et reconstituer ainsi des brins d'ADN correspondant à la totalité du gène codant pour la protéine de fusion. Une nouvelle réaction PCR à l'aide des oligonucléotides complémentaires des régions situées en amont et en aval du gène complet permet une amplification de ce dernier.
Les trois mutants construits par cette stratégie ont été validés par séquençage.

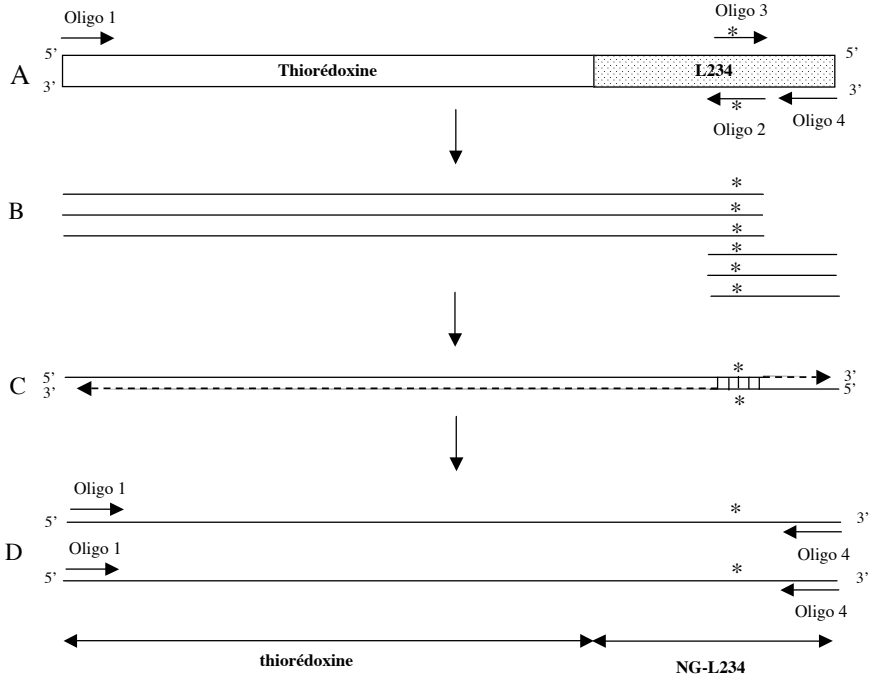


Figure 33 Stratégie utilisée pour la construction des mutants.
Le gène codant pour la thiorédoxine-L234 est amplifié en deux parties à l'aide de 4 oligonucléotides (A). Les oligonucléotides 1 et 4 sont complémentaires des régions amont et aval du gène tandis que les oligonucléotides 2 et 3 portent les mutations que l'on désire introduire Dans des conditions de température précise, les deux fragments purifiés (B) s'apparient ensuite entre-eux par une région complémentaire et la Taq polymérase peut synthétiser des brins d'ADN correspondant à la séquence complète du gène en prenant pour amorce la région commune appariée. Enfin, l'ajout d'oligonucléotides complémentaires aux régions amont et aval du gène complet (oligo 1 et 4) permet l'amplification de ce dernier (D).
Le sigle * correspond à l'emplacement de la base introduisant la mutation.

IV.2.2. Estimation de l'affinité pour le DEAE-sépharose Fast Flow des candidats tags mutants pré-purifiés

La première étape de l'analyse des mutants a consisté à vérifier que ces mutations ne diminuaient pas l'affinité spécifique du fragment L234 pour le DEAE-Sépharose. Dans cette optique, nous avons purifié les protéines thiorédoxine-L234, thiorédoxine-EDE-L234, thiorédoxine-V1V2V3-L2134 et thiorédoxine-NG-L234 sur colonne de chélation et avons testé leur affinité pour le DEAE de la façon suivante : chaque protéine a été déposée sur colonne DEAE-Sépharose Fast Flow puis, la colonne a été lavée avec un gradient linéaire de choline 0-800 mM. Si les mutations n'ont pas d'impact sur l'affinité du *tag*, nous nous attendons à ce que les protéines s'éluent à la même concentration en choline. Dans le cas contraire, les protéines de fusion doivent s'éluer à des concentrations différentes en choline. Le chromatogramme de l'expérience est présenté à la figure 34. Nous observons les résultats suivants :

- les résidus acide aspartique et acides glutamiques introduits dans le mutant thiorédoxine-EDE-L234 ne semblent pas avoir d'impact sur l'affinité pour le DEAE-Sépharose puisque cette protéine de fusion s'élue à même concentration en choline que la thiorédoxine-L234 (317 mM de choline, concentration calculée au départ du gradient théorique).
- la protéine de fusion mutante thiorédoxine-NG-L234 présente une moins bonne affinité spécifique pour le DEAE-Sépharose que la protéine non mutée (302 mM de choline, concentration calculée au départ du gradient théorique). Pour confirmer une diminution d'affinité spécifique du *tag*, il faudrait réaliser une chromatographie de la protéine thiorédoxine-NG-L234 pré-purifiée en utilisant d'abord un gradient de NaCl 0 - 2 M avant l'élution à la choline. Dans un tel cas, thiorédoxine-L234-NG-L234 devrait s'éluer dans le gradient NaCl au lieu de s'éluer spécifiquement en présence de choline. Il est possible que l'introduction concomitttante des deux mutations modifient légèrement la structure du *tag*, de telle sorte que la conformation des cages de résidus aromatiques constituant les sites de liaison soit également modifiée. L'introduction des mutations a aussi pour effet de modifier le point isoélectrique du fragment L234. Nous reviendrons sur ces points dans la discussion.
  Cependant, ce résultat n'exclut en rien que les mutations introduites stabilisent la structure globale de protéine de fusion lorsqu'elles sont introduites ensemble ou séparément. Néanmoins, dans le cadre de la création d'un *tag* de purification, il nous semble qu'une perte d'affinité spécifique pour le substrat est une raison suffisante pour écarter le candidat.

- contrairement aux deux premiers mutants, la protéine de fusion thiorédoxine-V1V2V3-L234 s'élue plus tard dans le gradient choline que la thiorédoxine-L234. Il semble donc que ce mutant ait une meilleure affinité spécifique pour le DEAE-Sépharose Fast Flow (353,6 mM choline, concentration calculée au départ du gradient théorique).

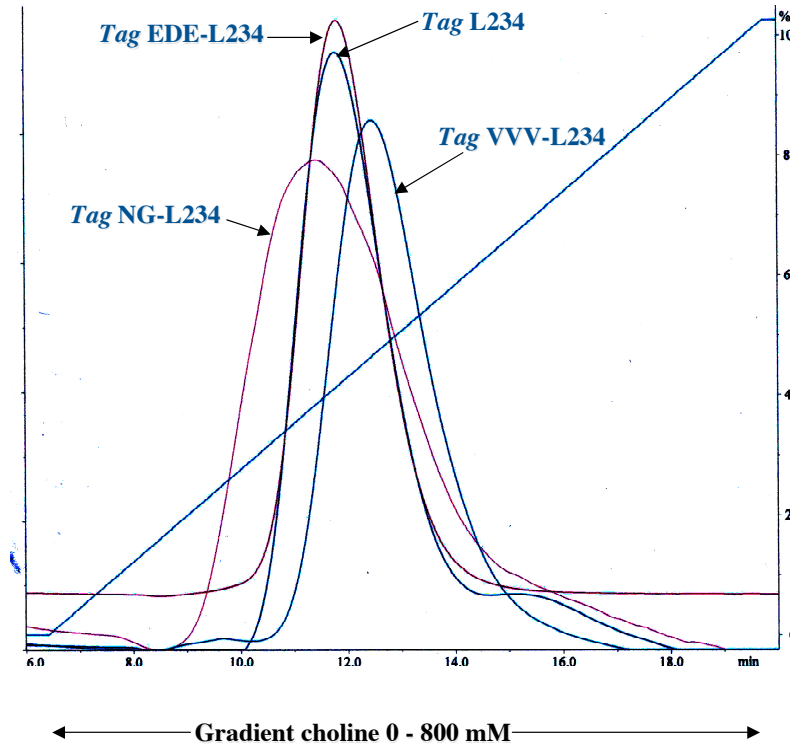


Figure 34 Profils chromatographiques des protéines de fusion thiorédoxine-L234, thiorédoxine-NG-L234, thiorédoxine-EDE-L234 et thiorédoxine-V1V2V3-L234 sur colonne DEAE-Sépharose Fast Flow après passage d'un gradient linéaire de choline 0-800 mM.

En conclusion, l'estimation de l'affinité pour le DEAE-Sépharose des *tag*s mutés montrent une augmentation de l'affinité spécifique du mutant V1V2V3-L234 pour son substrat lorsque la protéine de fusion est pré-purifiée sur colonne de chélation. Il semble donc que ces trois mutations stabilisantes définies par l'algorithme PoPMuSiC aient un effet indirect sur l'affinité spécifique globale de la protéine de fusion.
Par contre, le mutant NG-L234 présente une légère diminution d'affinité spécifique par rapport à celle du *tag* non muté et a donc été provisoirement écarté.

Les mutations définies sur base de résidus hydrophobes ne semblent pas avoir d'impact sur l'affinité spécifique pour le DEAE-Sépharose de la protéine de fusion pré-purifiée.

IV.2.3. Estimation de l'affinité pour le DEAE-Sépharose des protéines thiorédoxine-EDE-L234 et thiorédoxine-V1V2V3-L234 présentes dans la fraction soluble d'*E. coli.*

Comme nous venons de le voir au point précédent, il semble que l'introduction des mutations N et G diminue l'affinité spécifique du *tag* pour le DEAE ; les mutations EDE ne semblent pas avoir d'effet et l'introduction des mutations V1V2V3 semblent avoir un effet positif sur cette affinité. Les mesures ayant été faites lorsque les protéines de fusion sont prépurifiées, nous avons voulu savoir si ces effets étaient toujours observés lorsque les protéines de fusion étaient présentes dans une fraction soluble d'*E. coli*, c'est-à-dire lorsqu'elles étaient en compétition avec un grand nombre d'autres protéines pour le ligand.
Après surexpression de chaque protéine de fusion, les fractions solubles d'*E. coli* ont été déposées sur colonne DEAE-Sépharose. Un gradient NaCl 0-2 M a été appliqué, suivi d'un lavage phosphate 50 mM. Une élution subséquente à la choline 2% nous a permis de séparer les pics de protéines présentant une affinité spécifique pour le DEAE-Sépharose des pics de protéines adsorbées sur la colonne uniquement par des effets de charge.
Nos expériences montrent que la thiorédoxine-EDE-L234 s'élue soit dans le *flow through* (16% de la protéine), soit dans le gradient linéaire NaCl 0 - 2M (84% de la protéine). Cette expérience a été réalisée en triplicat.
Contrairement au candidat non muté thiorédoxine-L234, nous n'observons donc plus de pic protéique élué à la choline 2%. (figure 35)
Outre l'hypothèse d'un changement conformationnel suite à l'introduction des mutations, il faut envisager un effet de charge important lié au remplacement de trois résidus hydrophobes par trois résidus négatifs. Comme nous l'avons mentionné au point IV.1.1, ces mutations ont pour effet de faire diminuer le point isoélectrique du *tag* (tableau 26).

	IEP	ProtParam	MWCALC
Thiorédoxine-L234	6,5	6,03	6,81
Thiorédoxine-EDE-L234	5,97	5,64	6,31

Tableau 26 Calcul du point isoélectrique des protéines thiorédoxine-L234 et thiorédoxine-EDE-L234 par les programmes IEP (http://bioweb.pasteur.fr/seqanal/interfaces/iep.html), ProtParam (http://us.expasy.org/tools/protparam.html) et MWCALC (http://www.infobiogen.fr/services/analyseq/cgi-bin/mwcalc_in.pl).

Vu le principe de séparation des chromatographies par échange d'ions, il est envisageable que l'importance des interactions électrostatiques prennent le pas sur la spécificité du *tag* pour le DEAE-Sépharose. Nous reviendrons sur ce point dans la discussion.

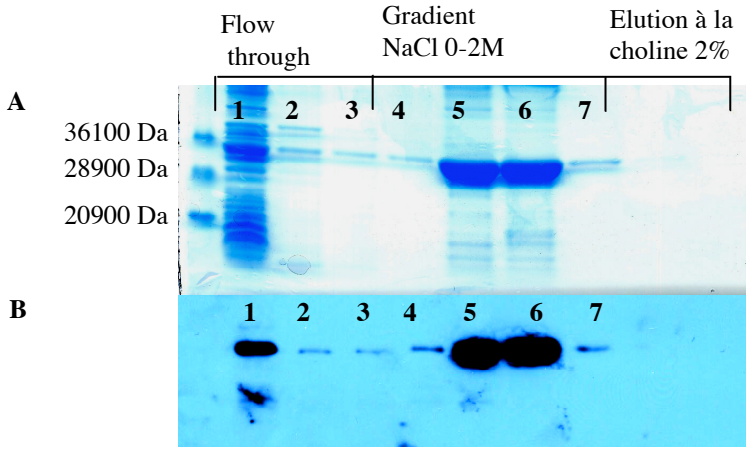


Figure 35 Estimation de l'affinité du mutant EDE-L234 pour le DEAE-Sépharose Fast Flow.
L'estimation a été réalisée par analyse en SDS-PAGE et Western Blot des fractions protéiques s'éluant de la colonne DEAE-Sépharose Fast flow après application de différents tampons.
A :analyse en SDS-PAGE après coloration au bleu de Coomasie
Pistes 1à 3 : fractions protéiques recueillies lors du lavage au phosphate 50 mM après dépôt de l'échantillon (*flow through*)
Pistes 4 à 7 : fractions protéiques recueillies lors du gradient NaCl 0-2M
Aucune fraction protéique n'a été recueillie lors du lavage à la choline 2%

B : analyse en Western Blot en utilisant un anticorps anti-*tag* 6 Histidines
Pistes 1à 3 : fractions protéiques recueillies lors du lavage au phosphate 50 mM après dépôt de l'échantillon (*flow through*)
Pistes 4 à 7 : fractions protéiques recueillies lors du gradient NaCl 0-2M
Poids moléculaire attendu du mutant EDE-L234 : 22700 Da

Par contre, trois essais de purification du candidat V1V2V3-L234 montrent que cette protéine de fusion s'élue en plus grande proportion en présence de choline 2% qu'en présence de NaCl (figure 36). Par densitométrie, nous estimons que 84% de la protéine s'élue à la choline alors que pour la thiorédoxine-L234, cette proportion n'était que de 45% (voir partie II). Un tel résultat suggère que l'affinité spécifique accrue pour le DEAE, observée lorsque la protéine de fusion est purifiée, est maintenue lorsque celle-ci est présente dans la fraction soluble d'*E. coli*. Nous obtenons donc une augmentation de l'affinité de ce candidat *tag* pour le DEAE-Sépharose suite à l'introduction de mutations stabilisantes.

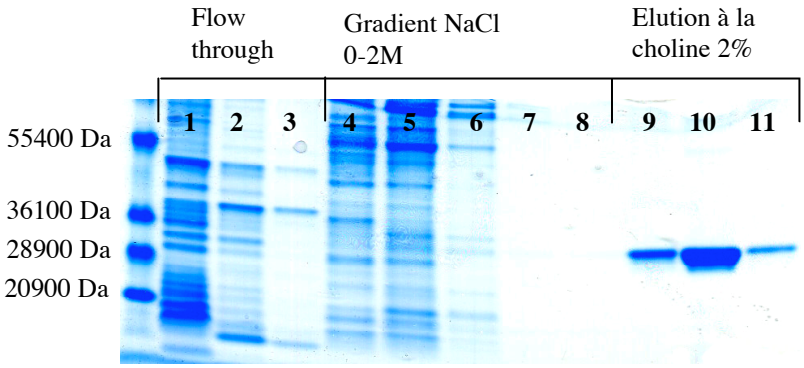


Figure 36 Estimation de l'affinité du mutant V1V2V3-L234 pour le DEAE-Sépharose Fast Flow.
L'estimation a été réalisée par analyse en SDS-PAGE des fractions protéiques s'éluant de la colonne DEAE-Sépharose Fast flow après application de différents tampons.
Pistes 1à 3 : fractions protéiques recueillies lors du lavage au phosphate 50 mM après dépôt de l'échantillon (*flow through*)
Pistes 4 à 8 : fractions protéiques recueillies lors du gradient NaCl 0-2M
Pistes 9 à 11 : fractions protéiques recueillies lors du lavage à la choline 2%
Poids moléculaire attendu du mutant EDE-L234 : 22746 Da

En conclusion, l'évaluation de l'affinité pour le DEAE-Sépharose du candidat *tag* EDE-L234 montre que, si ce *tag* conserve une affinité spécifique comparable au *tag* non muté lorsque la protéine de fusion est pré-purifiée sur colonne de chélation, les mutations introduites ne permettent plus à la protéine de s'éluer spécifiquement en présence de choline lorsque celle-ci est présente dans la fraction soluble d'*E. coli.*
Même lorsque la protéine de fusion est pré-purifiée, le candidat NG-L234 présente une diminution d'affinité spécifique pour le ligand par rapport à celle du *tag* L234. Ce candidat a donc été écarté.
Par contre, que la protéine thiorédoxine-V1V2V3-L234 soit pré-purifiée ou qu'elle soit présente dans la fraction soluble d'*E. coli*, le *tag* mutant présente une affinité spécifique accrue pour le DEAE que sa version non mutée. En conclusion, l'utilisation de l'algorithme PoPMuSiC pour définir des mutations stabilisantes est une voie intéressante puisqu'elle nous a permis de définir un *tag* mutant présentant une augmentation de son affinité spécifique pour le ligand, probablement par stabilisation de sa structure.

IV. 3. Estimation de la solubilité des protéines de fusion thiorédoxine-EDE-L234 et thiorédoxine-V1V2V3-L234

En parallèle aux expériences décrites ci-dessus, nous avons estimé la solubilité des protéines de fusion mutantes par deux procédés. Le but de notre travail consistant en la création d'un *tag* à vocation industrielle, une donnée importante est la quantité de la protéine d'intérêt qui peut être obtenue sous forme soluble. Nous avons donc utilisé comme première estimation de solubilité la quantité de protéines solubles produites par litre de culture. La surexpression des protéines de fusion a été réalisée dans les conditions mises au point au point II (IPTG 1 mM 21h à 18°C, au départ de cultures de 300 ml de LB + NaCl 0,3M). Les résultats sont repris au tableau 27.

Protéine de fusion	**Quantité (mg/l)**
Thiorédoxine-L234	64
Thiorédoxine-EDE-L234	92
Thiorédoxine-V1V2V3-L234	61

Tableau 27 Estimation de la quantité de protéine de fusion (en mg/l) produite dans 300 ml de culture après induction à l'IPTG 1 mM pendant 21h à 18°C.

Ces calculs, réalisés en triplicat, mettent en évidence qu'il y a en moyenne 30% de plus de protéine thiorédoxine-EDE-L234 que de protéine thiorédoxine-L234 ou thiorédoxine-V1V2V3-L234.

Dans un deuxième temps, nous avons calculé la concentration en sulfate d'ammonium nécessaire pour précipiter chaque candidat (phénomène de salting-in / salting-out lié à l'effet Hofmeister). Bien que cette expérience ne suffise pas à elle seule à prouver l'augmentation de solubilité d'une protéine, elle peut être utilisée en complément d'autres expériences (Mosavi and Peng, 2003) (Hachem et al., 1996). Les résultats sont repris à la figure 37.

Nous constatons que, comme pour le calcul de la quantité de protéine soluble obtenue, le mutant thiorédoxine-EDE-L234 se démarque des protéines de fusion thiorédoxine-L234 et thiorédoxine-V1V2V3-L234 puisqu'il faut 10% de sulfate d'ammonium en plus pour le précipiter à concentrations protéiques égales.

Ces résultats préliminaires suggèrent que les mutations des trois résidus hydrophobes en résidus hydrophiles ont une influence favorable sur la solubilité de la protéine de fusion thiorédoxine-L234.

Par contre, les mutations V1V2V3 ne semblent pas avoir d'influence sur la solubilité de la protéine de fusion puisque, dans nos expériences, le mutant se conduit comme la protéine non mutée.

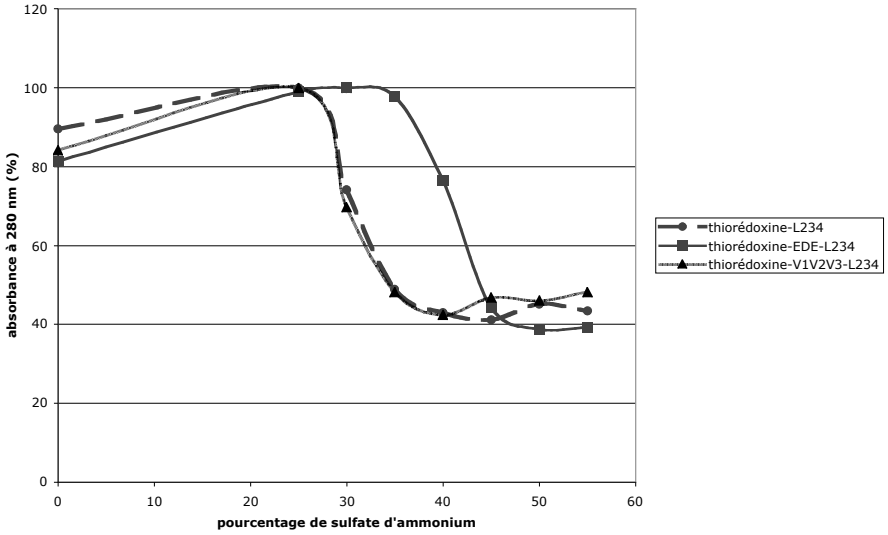


Figure 37 Détermination de la quantité de sulfate d'ammonium nécessaire pour précipiter 200 µg de thiorédoxine-L234, thiorédoxine-EDE-L234 et thiorédoxine-V1V2V3-L234 par mesure de l'absorbance à 280 nm. Cette mesure a été prise sur le surnageant après incubation des protéines en présence des différentes concentrations en sulfate d'ammonium pendant 12h à 4°C et centrifugation. Les valeurs d'absorbance observées pour chaque protéine et chaque concentration en sulfate d'ammonium ont été normalisée en ramenant la plus haute valeur d'absorbance à 100 % et en ajustant les autres valeurs à celle-là.

Ces données devraient être confirmées par d'autres types d'expériences. Nous pourrions notamment doser par densitométrie, après coloration au bleu de Coomasie, la proportion de protéine d'intérêt présente dans les fractions soluble et insoluble d'*E. coli*. Une augmentation de la quantité de thiorédoxine-*tag* muté dans la fraction soluble par rapport à celle obtenue pour la thioréoxine-*tag* non muté permettrait de renforcer l'idée d'une augmentation de solubilité du ou des mutants.

En conclusion, par calcul de l'accessibilité au solvant des résidus du domaine de liaison ClytA ou du fragment de domaine L234, nous avons mis en évidence trois résidus aromatiques potentiellement exposés au solvant dans le domaine tronqué et pouvant donc influer sur la solubilité du candidat *tag*. Ces résidus ont été mutés en acides aspartiques et glutamiques. Le calcul de la quantité de protéine soluble

obtenue ainsi que la détermination de la concentration en sulfate d'ammonium nécessaire pour précipiter respectivement les protéines thiorédoxine-L234, thiorédoxine-V1V2V3-L234 et la thiorédoxine-EDE-L234 montrent que l'introduction de ces dernières mutations semble augmenter la solubilité du *tag*. Ce résultat ne se vérifie pas pour le candidat *tag* V1V2V3-L234.

DISCUSSION GENERALE, CONCLUSIONS ET PERSPECTIVES

L'utilisation intensive de protéines purifiées dans divers secteurs économiques de notre société a pour conséquence un intérêt toujours croissant pour la mise au point de nouveau système de purification. Parmi les systèmes de purification par chromatographies, la chromatographie d'affinité possède un certain nombre d'avantages dont le principal est sa mise en œuvre simple, ne nécessitant pas l'isolement préalable et la caractérisation des propriétés biochimiques de la protéine d'intérêt. Généralement, elle consiste à fusionner un *tag* en N-terminal ou C-terminal d'une protéine cible. Ce dernier possède une affinité spécifique pour une matrice chromatographique sélectionnée. Un tel système permet assez souvent la purification de la protéine d'intérêt en une seule étape.
L'étude de systèmes biologiques faisant intervenir une interaction spécifique entre une protéine et son ligand est à la base de la conception de nombreux *tags*.
Dans ce contexte général, le but de notre travail était d'élaborer un *tag* de purification par affinité pour le DEAE-Sépharose. La conception du projet trouve son origine dans l'existence d'interactions spécifiques entre des protéines de surface de *Streptococcus pneumoniae* et les molécules de choline présentes dans la paroi de cette bactérie. La choline étant un analogue structural du DEAE, la base de notre projet a consisté à étudier le domaine de liaison à la choline de la protéine LytA pour en dériver un *tag* de purification.

1. Analyses bioinformatiques des domaines de liaison à la choline de protéines de surface de *Streptococcus pneumoniae* et de phages de streptocoques

Afin de mieux caractériser les résidus impliqués dans la reconnaissance de la choline, nous avons dans un premier temps défini un consensus de séquences des *repeats* constituant les domaines de liaison à la choline.
En effet, les seuls consensus connus au début de ce travail avaient été établis sur base d'alignements de séquences de *repeats* de domaine de liaison à la choline mais aussi de *repeats* d'autres types de domaines de liaison (voir introduction, paragraphe 2.5.2). Il n'existait donc pas de consensus de séquence propre aux *repeats* des domaines de liaison à la choline si ce n'est un consensus établi sur base de l'alignement des 12 *repeats* constituant les domaines de liaison à la choline des protéines LytA de *Streptococcus pneumoniae* et CPL1 du phage Cp-1 (Garcia et al, 1998).
Si nous comparons les consensus publiés au consensus de ce travail (tableau 28), nous constatons que le consensus que nous avons établi contribue à une

position	1	2	3	4	5	6
Consensus A {Giffard, 1994 #206}	X	souvent G	X	hydrophobe (1 à 3)	groupe de 4 à 6 résidus contenant souvant G et D et N	
Consensus B {Wren, 1991 #205}	T	G	W	x	T	I
Consensus C {Garcia, 1998 #52}	T	G	W	V	K/Q	D
Consensus D	pol/X	G/X	arom	phobe/X	ch/pol	X
structure	intéraction probable avec la paroi	zone de torsion	site de liaison	?	intéraction probable avec la paroi	?

position	7	8	9		10	11
Consensus A {Giffard, 1994 #206}					groupe de 1 à 4 résidus contenant des Y	
Consensus B {Wren, 1991 #205}	X	G	X	X	Y	Y
Consensus C {Garcia, 1998 #52}	N/K	G/D	T		W	Y
Consensus D	ch/pol	G	X		W	Y
structure	intéraction probable avec la paroi	zone de torsion	?		site de liaison	site de liaison

position	12	13	14	15	16	17
Consensus A {Giffard, 1994 #206}			groupe de 3 à 4 résidus variables			G
Consensus B {Wren, 1991 #205}	F		X	X	N	G
Consensus C {Garcia, 1998 #52}	Y	L	N/D	S	S/D	G
Consensus D	Y	phobe	ch/pol	petit	ch/pol	G
structure	pas évident	interaction hydrophobe avec le résidu 19	intéraction probable avec la paroi	zone de torsion	intéraction probable avec la paroi	zone de torsion

position	18		19	20
Consensus A {Giffard, 1994 #206}	2 résidus variables		1 résidu hydrophobe	
Consensus B {Wren, 1991 #205}	X	K	A	V
Consensus C {Garcia, 1998 #52}	A		M	A
Consensus D	petit		phobe	phobe
structure	zone de torsion		plancher du site de liaison	intéraction probable avec la paroi

Tableau 28 : Comparaison de consensus de séquences des motifs répétés composant les domaines de liaison à la choline. consensus A : (Giffard and Jacques, 199 consensus B : consensus défini par Wren (Wren, 1991), consensus C : consensus défini (Garcia et al., 1998), consensu D : consensus défini dans ce travail.

définition plus complète des classes de résidus situées de façon spécifique à certaines positions des *repeats* des domaines de liaison à la choline.
L'analyse ultérieure de la structure a confirmé l'importance de certaines classes de résidus à des positions particulières des *repeatS*. En effet, comme nous l'avons mentionné aux paragraphes I.4.1 et I.4.3, les résidus chargés et polaires, situés aux positions 1, 5, 7, 14 et 16, constituent des sillons reliant les sites de liaison à la choline. Ils pourraient établir des interactions électrostatiques et/ou des ponts hydrogènes avec les unités de glucan des acides téichoïques et lipotéichoïques de la paroi bactérienne. Ces interactions contribueraient à définir la grande affinité du domaine de liaison ClytA pour la paroi du pneumocoque (Fernandez-Tornero et al, 2001). Les positions 2, 8, 15, 17 et 18 sont principalement occupées par des glycines ou des petits résidus. Dans la structure, ils correspondent à des zones de torsion où des problèmes d'encombrements stériques peuvent se poser. Les résidus aromatiques aux positions 3, 10, 11 et le résidu hydrophobe en position 19 forment le site de liaison à la choline. Enfin, le résidu hydrophobe en position 13 établirait une liaison hydrophobe avec le résidu 19, contribuant probablement à stabiliser la structure.

La deuxième partie de l'approche bioinformatique consistait à rechercher par des méthodes de modélisation par homologie et de *threading* des structures connues potentiellement proches des structures des domaines de liaison à la choline. Ces analyses ne nous ont pas permis de sélectionner un candidat. La parution ultérieure de la structure de ClytA a confirmé l'impossibilité de sélectionner une structure pouvant servir de modèle puisque le domaine de liaison à la choline ClytA adoptait un *fold* qui n'avait jamais été répertorié dans les banques de données.

2. Sélection d'un fragment du domaine de liaison à la choline ClytA présentant toujours une affinité spécifique pour le DEAE-Sépharose

L'approche expérimentale que nous avons développée parallèlement aux analyses bioinformatiques a débuté par la sélection du plus petit fragment possible du domaine ClytA présentant toujours une affinité spécifique pour le DEAE-Sépharose, c'est-à-dire ne s'éluant de la colonne que par compétition avec la choline.
L'analyse de huit protéines de fusion nous a permis de sélectionner le fragment L234, composé de 71 résidus, et correspondant aux 9 derniers résidus du *repeat* 1, aux résidus des *repeats* 2, 3, 4 et aux 2 premiers résidus du *repeat* 5.

Des expériences de délétions progressives du domaine ClytA au départ de son extrémité carboxy-terminale avaient montré que le domaine devait conserver au moins les quatre *repeats* N-terminaux (89 résidus) pour présenter une affinité spécifique pour le DEAE (Garcia et al, 1994). Dans

leur expérience, Garcia et ses collaborateurs avaient surexprimé les différentes versions tronquées de LytA dans *E. coli* et déposé des extraits bactériens bruts sur des filtres de DEAE-cellulose. Les protéines d'intérêt avaient ensuite été éluées avec des tampons de force ionique croissante, suivis d'un lavage final des filtres à la choline 2%. Comme nous l'avons mentionné ci-dessus, la protéine comprenant encore 4 *repeats* s'éluait spécifiquement en choline alors que celle n'en possédant plus que trois (67 résidus) s'éluait totalement dans un tampon NaCl 0,75M. Nos expériences montrent qu'un fragment L234 (71 résidus) présent dans la fraction soluble d'*E. coli* présente encore une affinité spécifique pour le DEAE-Sépharose, puisqu'une partie de la protéine ne s'élue qu'en présence de choline. Ces résultats divergents démontrent l'importance, dans notre démarche, de réévaluer l'affinité de divers fragments du domaine de liaison ClytA, les conditions expérimentales dans lesquelles l'affinité pour le DEAE est mesurée étant différentes.

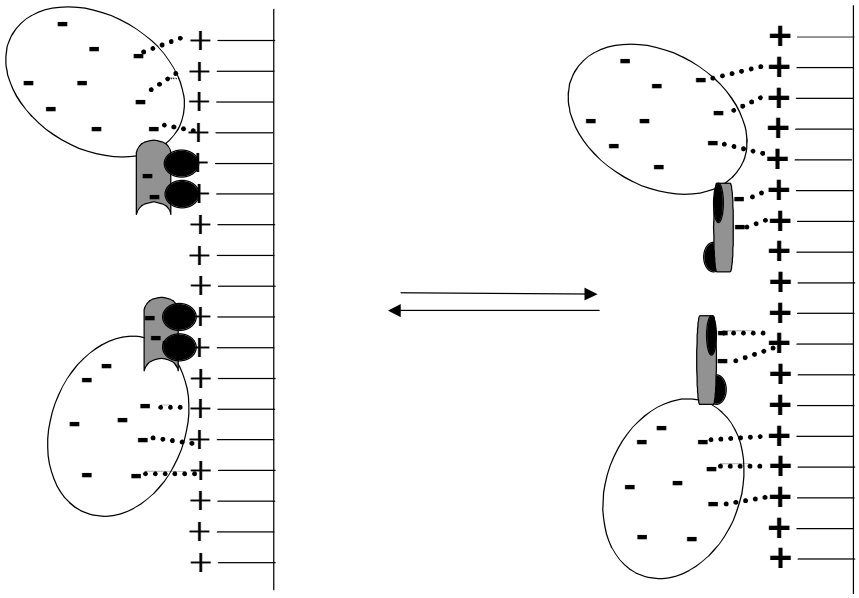
Cette expérience nous a aussi permis de montrer que tous les *repeats* composant le domaine de liaison à la choline ClytA ne semblaient pas avoir la même capacité à adopter une structure quand ils sont fusionnés en C-terminal de la thiorédoxine. En effet, la protéine de fusion thiorédoxine-L345 n'est pas produite. Outre l'adoption d'une structure très instable conduisant à sa dégradation dans la bactérie, il est possible que la fusion de ce fragment à la thiorédoxine donne un produit toxique pour *E. coli.* D'après nos expériences de fluorescence en présence d'un agent dénaturant, la thiorédoxine-L456, produite en très faible quantité sous forme soluble, ne semble pas capable de conserver une structure. L'adoption d'une structure très instable, conduisant à l'agrégation dans la bactérie, pourrait expliquer ce résultat.

L'estimation de l'affinité spécifique pour le DEAE de la thiorédoxine-L234 nous permet de dégager les observations suivantes :
Lorsque la thiorédoxine-L234 est pré-purifiée, elle ne s'élue de la colonne DEAE-Sépharose qu'en présence de choline. Elle présente donc une affinité spécifique pour le DEAE-Sépharose.
Par contre, lorsque cette protéine se trouve dans la fraction soluble d'*E. coli*, le fragment L234 ne permet plus de purifier la protéine d'intérêt dans sa totalité (de l'ordre de 50% de purification). Un résultat semblable est observé lorsque le fragment L234 est fusionné en aval de la protéine reporter MiaA.
En fraction soluble, il existerait donc deux populations de thiorédoxine-L234 que l'on ne discerne plus quand la protéine est prépurifiée.

Aucune de nos données expérimentales ni données de la littérature ne nous permet d'expliquer cette différence de comportement entre la protéine pré-purifiée et la protéine présente dans la fraction soluble bactérienne.

Quelques hypothèses peuvent cependant être avancées.

Premièrement, il est possible que le fragment L234 adopte une structure peu stable. En effet, bien que n'ayant pas réalisé de mesures précises de la stabilité de la thiorédoxine-L234 au pH, à la température ou aux agents dénaturants, les résultats de production obtenus pour les protéines thiorédoxine-L345 et thiorédoxine-L456 suggèrent un problème de stabilité au niveau de la structure adoptée par certains fragments du domaine de liaison ClytA, lorsque ceux-ci sont fusionnés à la thiorédoxine. De plus, dans la littérature, la présence de choline est citée comme un élément nécessaire au maintien de l'architecture en superhélice du domaine ClytA (Fernandez-Tornero et al, 2002). En fait, la choline protégerait du solvant les résidus hydrophobes de surface, formant les sites de liaison, et permettrait aux épingles à cheveux de rester empilées.
L'adoption d'une structure peu stable par le fragment L234 dans un milieu aqueux sans choline est donc envisageable. Dans ce contexte, il est possible que l'instabilité du *tag* se marque plus lorsque la protéine est présente dans la fraction soluble d'*E. coli* que lorsqu'elle est resuspendue dans un tampon phosphate 50 mM à pH 7,9, les conditions de salinité et de concentration protéique étant différentes.
L'instabilité conformationnelle pourrait notamment se traduire par une formation suboptimale de certains sites de liaison et aurait pour conséquence une diminution de l'affinité spécifique pour le ligand d'une partie de la population protéique. Cette population protéique s'adsorberait uniquement de manière aspécifique par l'établissement de liaisons électrostatiques entre ses charges négatives et le DEAE-Sépharose (figure 38).



A : thiorédoxine-L234 liée spécifiquement au DEAE-Sépharose par liaison du fragment L234 au DEAE

B : thiorédoxine-L234 adsorbée non spécifiquement au DEAE-Sépharose par des intercations électrostatiques

Figure 38 Schéma des équilibres entre une forme de la thiorédoxine-L234 liée spécifiquement au DEAE-Sépharose par le fragment L234 et une forme où elle est adsorbée de façon non spécifique sur la colonne par établissement d'interactions électrostatiques.
La thiorédoxine est représentée en blanc, le fragment L234 est coloré en gris et les deux sites de liaison à la choline sont en noir.
Le DEAE est symbolisé par le signe + et les interactions électrostatiques sont en pointillé.

Fragment L234 adoptant une conformation permettant la formation des sites de liaison à la choline

Fragment L234 déplié suite à un manque de stabilité de la structure, avec sites de liaison à la choline imparfaitement formés.

Deuxièmement, nous pourrions envisager l'interaction, *in vivo*, entre le fragment L234 et une protéine ou molécule de la bactérie. Si cette interaction masque les sites de liaison à la choline et qu'elle persiste en

partie lors de la préparation de la fraction soluble, la thiorédoxine-L234 ne peut plus s'adsorber sur la colonne que par liaisons électrostatiques.
Dans le cas d'une interaction avec une autre protéine, nous pourrions mettre en évidence une telle interaction en réalisant une analyse sur gel de polyacrylamide en conditions natives de la fraction soluble d'*E. coli* après surexpression de la thiorédoxine-L234, de la protéine de fusion prépurifiée sur colonne de chélation et des deux populations s'éluant de la colonne DEAE-Sépharose. Nous pourrions également envisager la mise en évidence d'une différence de poids moléculaire en utilisant un tamis moléculaire. D'autres analyses en gels de polyacrylamide à deux dimensions sont également envisageables.

Enfin, nous ne pouvons exclure que le *tag* L234 dimérise dans certaines conditions. En effet, des analyses d'ultracentrifugation montrent que le domaine ClytA dimérise lorsqu'il est en solution (Usobiaga et al, 1996). En parallèle, la détermination de la structure cristallographique de ce domaine a montré que l'interaction de deux monomères se réalise par appariement des épingles à cheveux des *repeats* 6 et des queues C-terminales (Fernandez-Tornero et al, 2001). Lors d'expériences cristallographiques ultérieures, un dimère de dimères du domaine ClytA a été mis en évidence (Fernandez-Tornero et al, 2002). Outre la dimérisation par appariement des extrémités C-terminales, une association des extrémités N-terminales est observée. Cette association interrompt l'extension normale de la superhélice en disruptant l'épingle à cheveux N-terminale des deux monomères et en formant un feuillet β antiparallèle étendu entre deux brins β de chaque monomère (fig 39).

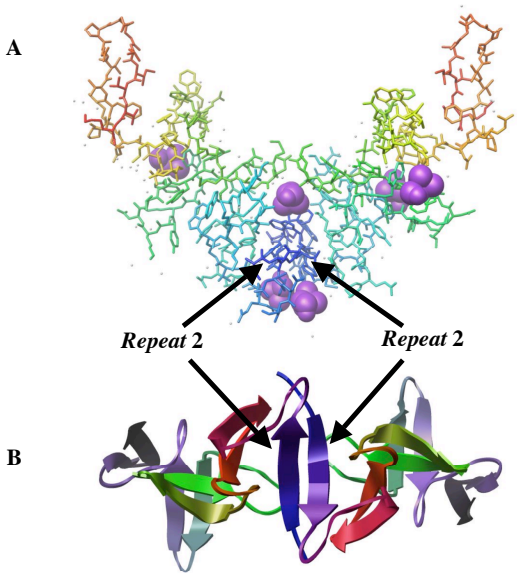


Figure 39 Dimérisation du domaine de liaison partiel ClytA par formation d'un feuillet β entre brins β de deux *repeats* 2.
A : vue latérale du dimère complet. Les molécules de choline sont symbolisées sous forme de boules
B : visualisation de l'appariement des *repeats* 2

Les auteurs concluent que le contact extensif disruptant la structure globale est probablement la conséquence du compactage du cristal. Cette dimérisation de dimère serait donc un phénomène fortuit lié aux conditions expérimentales particulières utilisées. Malgré cette conclusion, nous ne pouvons exclure que, lorsque la thiorédoxine-L234 est présente dans la fraction soluble d'*E. coli*, elle se trouve à des concentrations protéiques et dans des conditions de pH et salinité telles que ces dernières favorisent un état d'oligomérisation par formation d'un feuillet β entre les brins β des épingles à cheveux C-terminales.
Comme pour la deuxième hypothèse, une telle dimérisation aurait probablement pour effet de masquer certains sites de liaison à la choline, ayant de ce fait un effet sur l'affinité spécifique du *tag* L234 pour le DEAE-Sépharose.
Pour vérifier cette hypothèse, il faudrait également réaliser une analyse des différentes fractions sur gel de polyacrylamide en conditions natives ou sur tamis moléculaire, comme nous l'avons évoqué dans la deuxième hypothèse.

3. Analyse du *tag* mutant EDE-L234

Lorsque nous avons fusionné le fragment L234 en C-terminal des protéines reporter thiorédoxine et MiaA, nous avons observé qu'une grande proportion des protéines de fusion précipitaient sous forme de corps d'inclusion dans *E. coli*. Comme nous n'observions pas ce phénomène lorsque les protéines reporter étaient produites sans le *tag*, nous avons supposé que c'était la fusion du fragment L234 aux protéines reporter qui provoquaient l'insolubilisation de ces dernières.
Une cause possible de précipitation des protéines lors de leur production est l'exposition de zones hydrophobes à leur surface (Georgiou et Valax 1996) (King et al, 1996). L'insolubilisation de la thiorédoxine-L234 pourrait donc provenir de l'exposition de résidus hydrophobes au solvant suite au raccourcissement du domaine de liaison ClytA.
Nous pourrions aussi envisager que l'exposition de résidus hydrophobes en surface constitue une source d'instabilité structurale. En effet, Strub et ses collaborateurs (2004) ont montré que la mutation de résidus hydrophobes exposés au solvant en résidus hydrophiles constitue une stratégie efficace pour stabiliser les protéines.

Que les faibles quantités de thiorédoxine-L234 soluble obtenues soient liées à un problème de solubilité ou de stabilité, une stratégie pour augmenter la quantité de protéine soluble produite consiste à muter les résidus hydrophobes situés en surface en résidus hydrophiles (Strub et al, 2004)(Mosavi et al, 2003).

Par analyse visuelle de la structure CLytA, nous avons mis en évidence cinq résidus hydrophobes potentiellement exposés au solvant dans la structure tronquée. Trois de ces résidus ont été mutés en résidus hydrophiles chargés négativement.
Une première estimation de la solubilité des protéines thiorédoxine-L234 et thiorédoxine-EDE-L234 a été réalisée en calculant les quantités de protéines solubles obtenues. En parallèle, nous avons mesuré le pourcentage de sulfate d'ammonium nécessaire pour précipiter les deux protéines. Nos résultats montrent qu'il faut plus de sel pour précipiter la thiorédoxine-EDE-L234 que pour précipiter la protéine de fusion non mutante et que la quantité finale de protéines solubles est supérieure pour la thiorédoxine-EDE-L234 que pour la thiorédoxine-L234.
Bien que pour l'expérience au sulfate d'ammonium, l'introduction de trois charges négatives doive influencer la quantité de sel nécessaire à la précipitation de la protéine, ces deux expériences suggèrent que les mutations que nous avons introduites ont effectivement un effet sur la solubilité de la protéine de fusion.
Pour confirmer l'augmentation de solubilité du *tag* mutant, nous pourrions réaliser un dosage des fractions solubles et insolubles par densitométrie

après coloration d'un gel SDS-PAGE au bleu de Coomasie. Une augmentation de la protéine d'intérêt dans la fraction soluble par rapport à la fraction insoluble traduirait une augmentation de la production de protéine soluble et donc une augmentation probable de sa solubilité.
Une alternative consiste à étudier le temps de rétention des protéines d'intérêt sur colonne hydrophobe.
Pour confirmer les résultats obtenus avec le *tag* mutant EDE-L234, il faudrait également fusionner ce dernier en C-terminal d'autres protéines reporter afin de vérifier que les effets des mutations que nous observons ne sont pas la conséquence de la fusion particulière du *tag* à la thiorédoxine.

Lors d'essais de purification de la thiorédoxine-EDE-L234 sur colonne DEAE-Sépharose, au départ de la fraction soluble d'*E. coli*, on constate que la protéine s'élue dans le gradient NaCl. Nous n'observons donc plus d'élution spécifique à la choline. Une hypothèse permettant d'expliquer cette perte apparente d'affinité pour le DEAE-Sépharose est que les mutations introduites provoquent un changement structural ne permettant plus une formation optimale des sites de liaison.
Cependant, un autre facteur important est l'effet des trois charges négatives introduites qui diminuent le pI de la protéine (tableau 29).

	IEP	ProtParam	MWCALC
Thiorédoxine-L234	6,5	6,03	6,81
Thiorédoxine-EDE-L234	5,97	5,64	6,31
Thiorédoxine-NG-L234	6,68	6,18	6,98
Thiorédoxine-V1V2V3-L234	6,34	5,89	6,65

Tableau 29 Calcul du point isoélectrique des protéines thiorédoxine-L234, thiorédoxine-EDE-L234, thiorédoxine-NG-L234 et thiorédoxine-V1V2V3-L234 par les programmes IEP, ProtParam et MWCALC.

En effet, à pH 7,9, la thiorédoxine-EDE-L234 porte plus de charges négatives que la thiorédoxine-L234. Or, nous travaillons avec une matrice échangeuse d'anions qui permet à n'importe quelle protéine chargée négativement d'établir des interactions électrostatiques avec le DEAE et donc de s'adsorber sur la colonne. La thiorédoxine fusionnée à une version du *tag* L234 peut donc exister sous trois états (figure 38). Dans ce contexte, le fait d'introduire trois charges négatives supplémentaires dans le mutant thiorédoxine-EDE-L234 pourrait perturber l'équilibre entre ces trois états en les déplaçant vers l'état d'adsorption non spécifique. Il ne s'agirait donc pas d'une perte d'affinité spécifique du fragment L234 pour le DEAE mais plutôt d'un déplacement des équilibres établis entre la forme protéique liée spécifiquement, la forme en solution et la forme adsorbée par interactions

électrostatiques. Une donnée de la littérature vient étayer cette hypothèse. En effet, si nous partons du postulat que la structure adoptée par le fragment L234 est très proche de celle adoptée par ces *repeats* dans le domaine est complet, les charges introduites sont spacialement proches, créant une zone négative à la surface du *tag*. Or, des études sur la rétention de protéines sur colonnes échangeuses d'ions ont mis en évidence l'importance de la disposition des charges à la surface de la protéine (Kopaciewicz et al, 1983). Il semble que la distribution des charges à la surface des protéines soit un facteur plus déterminant que leur charge nette pour les purifier sur colonne échangeuse d'ions.
Dans ce contexte, une expérience intéressante consisterait à remplacer les résidus hydrophobes exposés en surface par des résidus polaires. Comme les résidus à muter possèdent une chaîne latérale assez volumineuse, il faudrait que les résidus les remplaçant aient également une chaîne latérale d'un volume comparable afin d'éviter la création de vides stériques. Il nous semble que l'asparagine et la glutamine sont de bons candidats. De nouvelles expériences de purification, par exemple avec un *tag* NQN, pourraient être réalisées. En utilisant un gradient de sel suivi d'une élution à la choline, ces expériences nous permettraient de vérifier si la perte apparente d'affinité observée avec le *tag* EDE-L234 se maintient (éventuel changement structural) ou si les résultats obtenus avec le *tag* EDE-L234 sont plutôt liés aux trois charges négatives introduites.

4. Introduction de mutations potentiellement stabilisantes dans la séquence du *tag* L234

Comme nous l'avons mentionné précédemment, l'existence de deux populations protéiques présentant un comportement différent sur DEAE-Sépharose pourrait être expliquée par l'instabilité de la structure adoptée par le fragment L234. Les faibles quantités de thiorédoxine-L345 et thiorédoxine-L456 obtenues ainsi que certaines données de la littérature mettant en évidence l'importance de la choline dans la stabilisation de la structure de ClytA nous renforcent dans l'idée que le facteur stabilité peut être important et qu'il faut en tenir compte dans un contexte d'amélioration du *tag* L234. Partant de ce postulat, nous avons comparé la séquence du fragment L234 au consensus de *repeats* afin de mettre en évidence des résidus statistiquement sous-ou sur-représentés à certaines positions des *repeats*. En parallèle, nous avons réalisé une analyse de cette séquence par l'algorithme PoPMuSiC afin de définir des mutations potentiellement stabilisantes. Ces analyses ont abouti à la définition de deux groupes de mutations.

4.1. Analyse du *tag* mutant NG-L234

Les résidus Alanine 245 et acide aspartique 246 du *repeat* 4 ont été mutés respectivement en asparagine et en glycine.
Avant d'analyser les résultats que nous avons obtenus, il est cependant important de rappeler que l'algorithme PoPMuSiC a été conçu pour définir des mutations stabilisantes individuelles. En d'autres termes, il ne garantit en rien un effet additif des mutations sur la stabilité de la protéine. Nous avons cependant décidé de grouper ces deux mutations car, lors de l'alignement de 125 *repeats* provenant de 19 domaines de liaison à la choline, le couple NG apparaît 46 fois sur 125 aux positions 7 et 8 des *repeats*.
Ces mutations augmentent légèrement le pI de la protéine suite au remplacement d'une charge négative (tableau 29). Une compétition entre l'effet des charges et l'affinité spécifique pour le DEAE nous semble moins probable que dans le cas du mutant EDE-L234. En effet, pour ce mutant, nous supprimons une charge négative, rendant donc la protéine plus positive que sa version non mutée à pH 7,9. Si ce changement de charge doit avoir une influence, ce serait dans le sens d'une répulsion entre la protéine et la matrice, avec pour conséquence un déplacement éventuel des équilibres vers l'état d'adsorption spécifique. De plus, nous ne créons pas de zone particulièrement chargée en surface de la protéine.
Cependant, lorsque la thiorédoxine-NG-L234 est prépurifiée, elle présente une légère perte d'affinité spécifique pour le DEAE-Sépharose.
Comme pour le mutant EDE-L234 nous ne pouvons pas non plus écarter l'hypothèse d'un changement structural suite à l'introduction des mutations, avec un éventuelle conséquence sur la formation optimale d'un ou des deux sites de liaison.
Ce résultat préliminaire nous a amené à écarter le candidat NG-L234 dans la suite de nos analyses. Avant d'éliminer définitivement ce *tag*, il faudrait refaire une chromatographie en utilisant un gradient de sel suivi d'une élution à la choline afin de vérifier l'affinité de la protéine mutante dans la fraction soluble de *E. coli*. Une autre expérience à réaliser consiste à fusionner le *tag* NG-L234 derrière d'autres protéines reporter afin de vérifier que la légère perte d'affinité pour le DEAE, observée pour la thiorédoxine-NG-L234, ne résulte pas d'interactions particulières entre le *tag* mutant et la protéine reporter choisie mais que ce résultat se généralise quelle que soit la protéine reporter sélectionnée.

De plus, le critère final utilisé pour sélectionner les mutants est le comportement de la protéine de fusion sur colonne DEAE-Sépharose. Il faut cependant garder à l'esprit que nous avons défini des mutations stabilisantes en postulant que ces dernières allaient peut-être avoir un effet sur l'affinité. Le test d'affinité que nous avons effectué n'exclut donc en rien que les mutations introduites soient effectivement stabilisantes pour la structure,

bien qu'elles aient un effet négatif apparent sur l'affinité spécifique du *tag* dans les conditions testées. Il faudrait réaliser des tests de dénaturation en urée tels que ceux décrits dans Dumoulin et al (2002) ou dans Gilis et al (2003) ou estimer la stabilité thermique de la protéine pour valider ou écarter définitivement ces mutations en terme de mutations stabilisantes.
Enfin, il faudrait également réintroduire les mutations une à une, tester la stabilité aux agents dénaturants des protéines mutantes obtenues ainsi que leur affinité pour le DEAE-Sépharose.

4.2. Analyse du *tag* mutant V1V2V3-L234

Les résidus chargés situés en positions 4 des *repeats* 2, 3 et 4 du fragment L234 ont été remplacés par des valines. Comme pour le mutant NG-L234, ces mutations ont été sélectionnées par l'algorithme PoPMuSiC et proposée, en parallèle, suite à l'analyse de la séquence des *repeats* 2, 3 et 4 et leur comparaison au consensus de séquence. Les mutations augmentent le pI de la protéine suite à la perte de deux charges positives et une charge négative (tableau 29). Comme nous l'avons postulé pour le mutant thiorédoxine-NG-L234, il nous semble qu'une augmentation du pI, aussi petite soit-elle, ne devrait théoriquement avoir qu'un effet positif sur le rendement de purification, un déplacement d'équilibre entre la proportion de protéines liées spécifiquement et la proportion de protéines adsorbées de manière aspécifique devant plutôt favoriser l'établissement d'interactions spécifiques.
En pratique, que la thiorédoxine-V1V2V3-L234 soit prépurifiée ou présente dans la fraction soluble *d'E. coli*, la proportion de protéine s'éluant spécifiquement en choline est accrue. Il semble donc que nous ayons un gain d'affinité spécifique pour le DEAE-Sépharose.

Pour continuer la caractérisation de ce candidat *tag*, nous devrions valider les mutations V1V2V3 en tant que mutations stabilisantes. Cette démarche nécessite des mesures de stabilité *in vitro* en regardant soit l'équilibre de dénaturation en présence d'agents dénaturants soit la stabilité thermique. Si le gain de stabilité s'avère réel, nous pourrions enfin confirmer une de nos hypothèses, à savoir qu'il existe un lien entre la stabilité de la structure adoptée par le fragment L234, la conformation optimale des sites de liaison à la choline et l'affinité spécifique observée sur colonne.
Enfin, nous devrions également fusionner le *tag* V1V2V3 derrière d'autres protéines reporters afin de vérifier que l'affinité pour le DEAE observée avec la thiorédoxine-V1V2V3-L234 reste inchangée lorsque le *tag* mutant est fusionné derrière d'autres protéines d'intérêt.

CONCLUSION ET PERSPECTIVES

Au terme de ce travail, nous avons sélectionné un fragment du domaine de liaison ClytA constitué des *repeats* 2, 3 et 4. Ce fragment permet de purifier les protéines thiorédoxine et MiaA sur colonne DEAE-Sépharose lorsque ces protéines sont présentes dans la fraction soluble d'*E. coli*. Bien qu'une partie des protéines d'intérêt s'éluent lors du gradient de sel, la fraction protéique s'éluant en choline semble purifiée correctement puisque nous n'observons pas de bandes contaminantes au niveau de cette fraction dans un gel SDS-page coloré au bleu de Coomasie. Cependant, le *tag* L234 semble provoquer l'insolubilisation des protéines de fusion lors de leur production dans la bactérie.

En vue d'améliorer les propriétés physico-chimiques du fragment L234, nous avons défini plusieurs stratégies.

D'une part, la faible solubilité du *tag* pourrait être due à l'exposition de résidus hydrophobes suite au raccourcissement du domaine de liaison ClytA. Ces résidus hydrophobes exposés en surface pourraient également contribuer à déstabiliser la structure adoptée.

D'autre part, une instabilité structurale, liée ou non à la présence de résidus hydrophobes en surface de la protéine, pourrait avoir un effet sur la solubilité du *tag*. Cette instabilité présumée pourrait aussi entraîner une formation suboptimale des sites de liaison à la choline. Ces hypothèses permettraient d'expliquer pourquoi une partie des protéines de fusion s'élue dans le gradient NaCl. En résumé, nous postulons qu'il existe un lien d'une part entre la stabilité de la structure et sa solubilité et d'autre part entre la stabilité structurale et l'affinité spécifique pour le DEAE-Sépharose.

Partant de ces hypothèses, trois résidus hydrophobes potentiellement exposés en surface du *tag* ont été remplacés par des résidus acides. Ces mutations ont permis d'obtenir une plus grande quantité de la protéine d'intérêt sous forme soluble après surexpression dans *E. coli*. Il semble donc que les résidus hydrophobes mutés soient des éléments critiques intervenant dans la solubilité des protéines de fusion. Les mutations que nous avons définies entraînent cependant une perte apparente d'affinité pour le ligand. Celle-ci peut être due à des modifications structurales modifiant la géométrie des sites de liaison à la choline ou à une influence importante des charges négatives introduites.

En parallèle, nous avons défini, par l'algorithme PoPMuSiC, cinq mutations stabilisantes. Trois de ces mutations se sont révélées intéressantes puisqu'elles nous ont permis d'augmenter le rendement de purification sur DEAE-Sépharose. Il y a donc augmentation de l'affinité spécifique apparente pour le DEAE de la thiorédoxine-L234. Bien que n'ayant pas confirmé la stabilisation de la structure par des mesures *in vitro*, il semble donc qu'un lien entre la stabilité structurale et l'affinité spécifique du fragment pour le DEAE puisse être évoqué. Ces mutations n'ont cependant

pas d'effet sur la solubilité des protéines de fusion, les quantités de protéines solubles obtenues restant inchangées.

En perspective de ce travail, plusieurs pistes peuvent être exploitées pour continuer à améliorer le *tag*.

Nous pourrions créer une série de *tags* mutants, en remplaçant les trois résidus hydrophobes par une série de résidus hydrophiles, tout en conservant les mutations V1, V2 et V3.
Cette approche semi-aléatoire nécessiterait la mise au point de nouveaux systèmes d'analyse, vu le nombre potentiellement élevé de candidats.
Un système d'analyse en plaques multipuits est envisageable.
En testant l'aptitude des nouveaux *tags* mutants à purifier une protéine d'intérêt sur DEAE et en calculant, en parallèle le taux de protéines recombinantes solubles produites par *E. coli*, nous pourrions sélectionner ainsi des *tags* plus solubles permettant une purification protéique sur DEAE-Sépharose.

Un autre système d'analyse consisterait d'abord à sélectionner des *tags* n'entraînant pas de perte de solubilité de la protéine de fusion puis à tester l'affinité de ces *tags* sélectionnés pour le DEAE-Sépharose. Nous pourrions, par exemple, fusionner les candidats *tags* derrière le gène codant pour la chloramphénicol-acétyl-transférase (CAT), responsable de la résistance de la bactérie au chloramphénicol (Petrounia et al, 2000). Si le candidat *tag* provoque l'insolubilisation de la CAT, celle-ci se retrouvera emprisonnée dans des corps d'inclusion et ne permettra pas à la bactérie de se développer en présence de concentrations élevées en chloramphénicol. De cette façon, seuls les *tags* solubles seront sélectionnés puis leur affinité pour le DEAE sera testée. D'autres systèmes reporters de la solubilité de protéines surexprimées utilisent la Green Fluorescent Protein (GFP) ou un domaine de la β-galactosidase (Pedelacq et al, 2002) (Wigley et al, 2001).

Outre le test d'affinité sur DEAE-Sépharose, nous pourrions envisager de mesurer la liaison de choline marquée (tritium) pour estimer l'affinité spécifique des candidats *tags*.

Enfin, nous ne pouvons terminer ce travail sans mentionner une autre technique très utilisée dans la conception de nouveaux *tag* de purification par affinité (Thomas et al, 1993) (Schatz, 1993). En effet, si nous n'avions pas eu de contraintes de licence d'exploitation liées à l'utilisation de technologies brevetées, nous aurions pu aborder la problématique de ce travail en utilisant la technique de présentation de peptides en surface de phages (*phage-display*) ou de bactéries (*bacterial display*). La technique du *phage-display* consiste à cloner une banque de peptides aléatoires dans le gène codant soit pour la protéine de surface PIII soit pour la protéine de

surface PVIII (Smith, 1985). Au cours du cycle de vie, les phages nouvellement produits présentent donc à leur surface une banque de peptides ou protéines. En incubant ces phages avec un ligand particulier puis en lavant dans des conditions de stringence telles que seuls les phages présentant une affinité particulière pour le substrat, via le peptide qu'ils expriment en surface, restent adsorbés, on sélectionne les peptides possédant l'affinité recherchée. Le principe du bacterial display est semblable (Lu et al, 1995). Dans ce cas, une banque de dodécamères aléatoires est clonée dans un domaine non essentiel de la principale protéine flagellaire bactérienne, FliC. Le principe de sélection de peptides présentant une affinité particulière pour substrat reste le même que celui appliqué pour les phages.

L'utilisation de banques de peptides ou protéines en surface de phages ou de bactéries présentent plusieurs avantages majeurs :
Premièrement, tous les variants de la banque sont théoriquement soumis à une même étape de sélection conduisant directement à la sélection de quelques candidats. Cette approche permet donc de tester un beaucoup plus grand nombre de candidats que dans une approche plus classique où les différents candidats sont testés individuellement.
Deuxièmement, le fait que le peptide exprimé en surface du phage ou de la bactérie soit codé par un insert incorporé dans son génome permet de créer un lien physique direct entre le phénotype (le peptide présentant l'affinité recherchée) et le génotype (la séquence d'ADN codant pour ce peptide).
Enfin, dans le cadre de la création d'un *tag*, cette approche permettrait probablement de sélectionner directement des petits *tags* constitués seulement de quelques résidus.

.

MATERIEL ET METHODES

Cette partie reprend les souches, plasmides et oligonucléotides utilisés durant ce travail et développe les méthodes spécifiques employées.

1. Souches bactériennes et vecteurs plasmidiques

1.1. Les souches d'*Escherichia coli*

Les souches utilisées sont cultivées dans du milieu Luria-Bertani (LB) (par litre : 10g de tryptone, 5g de NaCl, 5g de Yeast Extract) et maintenues sur du LB agar 1,5%.

1.1.1. Escherichia coli *DH10B*

La souche DH10B (Gibco BRL) a été utilisée pour toutes les constructions plasmidiques de ce travail.
Son génotype complet est F- *mcrA* D(*mrr-hsdRMS-mcrBC*) f 80 *lac*ZDM15 D*lacX74 deoR recA1 araD139* D(*ara leu*)7697 *galU galK l - rpsL endA1 nupG*
Quelques caractéristiques importantes de ce génotype sont :
- *mcrA* D(*mrr-hsdRMS-mcrBC*) : élimination des systèmes de restriction mcrA,
mcrBC, mrr et hsdRMS
- *recA*1 : réduit la recombinaison au millième de son taux normal, minimisant ainsi la recombinaison entre l'ADN endogène et exogène
- *endA*1 : abolition de l'activité de l'endonucléase I non spécifique. Les souches portant cette caractéristique permettent des préparations d'ADN de meilleure qualité.
- *lac*ZDM15 : sélection de clones contenant un plasmide recombinant par un test bleu-blanc. Ce test est basé sur le phénomène de la complémentation alpha où une partie du gène codant pour la β-galactosidase est délétée dans le génome bactérien et cette même portion de gène (fragment lacZ') est portée par un plasmide. Lors de l'introduction d'un tel plasmide dans les bactéries DH10B, les protéines partielles codées par les deux fragments de gène s'associent pour former une β-galactosidase fonctionnelle. Si l'on ajoute dans le milieu un substrat chormogène incolore tel que le X-gal, celui-ci est capable d'être clivé en un colorant bleu par l'enzyme reconstituée. Si un fragment d'ADN est cloné sur le plasmide dans le gène codant pour le fragment lacZ', il n'y a plus production d'une β-galactosidase fonctionnelle et le substrat X-gal n'est plus clivé. Les colonies bactériennes porteuses d'un plasmide recombinant seront donc blanches.

1.1.2. Escherichi coli *BL21(λDE3) (Novagen)*

Cette souche bactérienne a été utilisée pour les surexpressions de toutes les protéines de fusion.
Elle est naturellement déficiente en protéase Lon.
Son génotype complet est F- *ompT* $hsdS_B$ (r_B- m_B-) *gal dcm* (λDE3)

Ses principales caractéristiques sont :
- *OmpT* : déficience en protéase OmpT, permettant d'obtenir de ce fait des rendements plus importants en protéines recombinantes intactes
- $hsdS_B$ (r_B- m_B-) : mutation dans le système restriction/modification de la bactérie, permettant le clonage d'ADN exogène sans clivage de ce dernier par des endonucléases de restriction endogènes
- (λDE3) : bactériophage dérivé du phage lambda, portant le gène codant pour l'ARN polymérase T7 sous le contrôle du promoteur *lacUV*5, inductible à l'IPTG

1.2. Vecteurs plasmidiques

1.2.1. Vecteur pBluescript SK(+) (Stratagene)

Le plasmide pBluescript SK(+) a été utilisé pour le clonage intermédiaire de tous les produits PCR.
Ce phagmide de 2,964 kb possède les caractéristiques suivantes :
- une origine de réplication chez *Escherichia coli* (OriC), dérivée du plasmide pUC
- un gène de résistance à l'ampicilline
- la portion de gène *lacZ*' codant pour une partie de la b-galactosidase. *LacZ*' est sous le contrôle du promoteur *lac* et porte un site multiple de clonage. Ce système permet la sélection de plasmides recombinants par un test bleu-blanc.

1.2.2. Vecteur pET15b (Novagen)

Ce vecteur est dédicacé à la surexpression de protéines dans les bactéries BL21(λDE3)
Il est caractérisé par :
- un site multiple de clonage précédé d'une séquence codant pour le *tag* 6 Histidines, le tout sous le contrôle du promoteur T7. Entre le *tag* 6 Histidines et le site multiple de clonage se trouve un site de clivage pour la thrombine, permettant d'éliminer le *tag* après production de la protéine dans la bactérie.
- un gène de résistance à l'ampicilline
- une origine de réplication chez *Escherichia coli* , dérivée du plasmide pBR322
- le gène *lac*I codant pour le répresseur lac. En absence de β-galactosides, ce répresseur se lie à l'opérateur *lac*O situé sur le promoteur T7, empêchant la transcription du gène situé en aval de ce promoteur.

2. Amplification et clonage de la thiorédoxine et des fragments de domaine ClytA dans le vecteur pET15b

De façon générale, la thiorédoxine et les fragments de domaine de liaison ClytA ont été amplifiés au départ de plasmides porteurs du gène d'intérêt ou d'ADN génomique de *Streptococcus pneumoniae* par la réaction de "Polymerase Chain Reaction" (PCR).
Nous avons utilisé des amorces porteuses de sites de restriction permettant le clonage ultérieur de l'ADN amplifié dans le vecteur pET15b.
La thiorédoxine a été amplifiée au départ d'un plasmide porteur du gène *trxA* d'*Escherichia coli* et a été clonée aux sites de restriction *Nde*I et *Bam*HI du vecteur pET15b.
Les différentes versions tronquées du domaine de liaison ClytA ont été amplifiées par PCR au départ du plasmide pCUZ1 (Sanchez-Puelles et al., 1992) ou d'ADN génomique de *Streptococcus pneumoniae* en utilisant les oligonucléotides repris au tableau 30. Chaque fragment amplifié a été cloné en aval de la thiorédoxine aux sites de restriction *Asp*718 et *Sac*I du plasmide pET15b-thiorédoxine.

NOM	UTILISATION	SEQUENCE	CARACTERISTIQUES
oligo thio2AM	amplification de la thiorédoxine	5'-GGC**CATATG**AGCGATAAAATTATTCAC-3'	site *Nd* eI
oligo thio aval	amplification de la thiorédoxine	5'-TTCCTCGACGCTAACCTGGCG**GGTACCGTCGA** CAAAGAGCTCCCCGGGTGAATGACTGAGGATCCAAGC-3'	sites *Asp* 718, *Sal* I, *Sac* I *Sma* I et *Bam* HI
oligo CLytA amont	amplification des fragments de domaine L2, L23, L234, L2345, L23456	5'-ATG**GGTACC**GTACATTCAGACGGCTCTTAT-3'	site *Asp* 718
oligo Clyt aval 2	amplification des fragments de domaine CLytA, L23456, L456	5'-CTGGCAGACAGGCCAGAA*TGA* **GAGCTC**ATC-3'	site *Sac* I un stop avant le site de restriction
oligo repeat 1	amplification des fragments de domaine CLytA, L123, L1234	5'-GATATG**GGTACC**TTGACGATTGAAACAGGCTGGCA GAAGAATGACACTGGCTACTGGTACGTACATTCAGACGGCTCTTAT-3'	site *Asp* 718
oligo 122 rev	amplification du fragment de domaines L2	5'-TATATGCTTGCAGACCGC*TGA* **GAGCTC**ATC-3'	site *Sac* I un stop avant le site de restriction
oligo 123AV	amplification du fragment de domaines L23	5'-GACAACTCAGGCGAAATGGCTTGAGGATCCTATGGC-3'	site *Bam* H I un stop avant le site de restriction
oligo 123 rev ou REP3AV	amplification du fragment de domaines L23	5'-AACTCAGGCGAAATGGCT*TGA* **GAGCTC**ATC-3'	site *Sac* I un stop avant le site de restriction
oligo 124 rev	amplification du fragment de domaines L234	5'-GGTGCCATGAAGACAGGC*TGA* **GAGCTC**ATC-3'	site *Sac* I un stop avant le site de restriction
oligo 345 AM	amplification du fragment de domaine L345	5'-GCGATT**GGTACC**GGCTATATGCTTGCAGAC-3'	site *Asp* 718
oligo 456 AM	amplification du fragment de domaine L456	5'-GCGATT**GGTACC**GGCGAAATGGCTACAGGC-3'	site *Asp* 718
oligo 345 AV	amplification du fragment de domaine L345	5'-TTAGACGCTAAAGAAGGC**GAGCTC**GTCGAG-3'	site *Sac* I
oligo 456 AV	amplification du fragment de domaine L456	5'-TACTACCTCAAACCAGAC**GAGCTC**GTCGAG-3'	site *Sac* I

Tableau 30 Liste des oligonucléotides utilisés pour amplifier par PCR la thiorédoxine et les fragments de domaine ClytA.

Comme expliqué à la figure 33 des résultats expérimentaux, la construction des versions mutantes du *tag* L234 ont été réalisée par réactions PCR successives en utilisant des oligonucléotides porteurs des mutations désirées (tableau 31). Les fragments codant pour les protéines de fusion mutées ont ensuite été clonées aux sites *Nde*I et *BamH*I du vecteur pET15b.

NOM	SEQUENCE	CARACTERISTIQUES DE L'OLIGONUCLEOTIDE
T3AM	5'-GCACCTGTGGCGCCGGTG-3'	amplification de la protéine de fusion complète thiorédoxine-tag
MUTAV	5'-ACAGCTTATCATCGATAAGCTAG-3'	amplification de la protéine de fusion complète thiorédoxine-tag
VAL1forward	5'-GACAAGTTTG**T**GAAAATCAATGGC-3'	Introduction de la mutation V1
VAL1 reverse	5'-GCCATTGATTTTC**A**CAAACTTGTC-3'	Introduction de la mutation V1
VAL2forward	5'-GACCGCTGG**GT**GAAGCACACAGAC-3'	Introduction de la mutation V2
VAL2reverse	5'-GTCTGTGTGCTTC**AC**CCAGCGGTC-3'	Introduction de la mutation V2
VAL3forward	5'-GACCGCTGG**GT**GAAGCACACAGAC-3'	Introduction de la mutation V3
VAL3reverse	5'-GTCTGTGTGCTTC**AC**CCAGCGGTC-3'	Introduction de la mutation V3
LOOPforward	5'-GCTACAGGCTGGAAGAAAATC**AA** T**G**GTAAGTGGTACTATTTCAAC-3'	Introduction des mutations NG
LOOP reverse	5'-**AC**CA**TT**GATTTTCTTCCAGCCTGTAGC-3'	Introduction des mutations NG
PHOBEforward	5'-**GAA**AAGAAAATCGCTGATAAG**GAT**TAC**G**A**A** TTCAACGAAGAAGGTG-3'	Introduction des mutations EDE
PHOBEreverse	5'-**T**T**C**GTA**ATC**CTTATCAGCGATTTTCTT**TTC** GCCTGTAGCCATTTCGCCTG-3'	Introduction des mutations EDE

Tableau 31 Liste des oligonucléotides utilisés pour introduire les mutations d'intérêt dans la séquence codante du *tag* L234.
Les bases introduisant les mutations sont indiquées en gras et soulignées.

4. Surexpression de la thiorédoxine et des protéines de fusion thiorédoxine-domaine tronqué

4.1. Principe de la surexpression de protéines recombinantes avec le système pET

Le système pET est un des systèmes les plus performants actuellement développé pour le clonage et l'expression de protéines recombinantes dans *E. coli*.
La bactérie BL21(λDE3) contient, dans son génome, le gène codant pour l'ARN polymérase du phage T7 sous le contrôle du promoteur *lacUV*. En absence de b-galactosides, la bactérie synthétise de façon constitutive un répresseur lac se liant à l'opérateur du promoteur *lacUV*5, empêchant ainsi toute transcription du gène codant pour l'ARN polymérase du phage T7. Lorsqu'un b-galactoside est ajouté dans le milieu, il se lie au répresseur, empêchant ce dernier de se lier à l'opérateur. et permet ainsi la production d'ARN polymérase du phage T7. Si la bactérie est transformée par un plasmide porteur du promoteur du phage T7, l'ARN polymérase phagienne est tellement active et sélective qu'elle détourne une grande partie des ressources cellulaires pour produire la protéine codée par le gène sous le contrôle du promoteur T7.
En pratique, on utilise un β-galactoside de synthèse, l'isopropyl-β-D-thiogalactoside (IPTG).

4.2. Conditions standards de surexpression de protéines recombinantes avec le système pET et préparation des fractions soluble et insoluble

- ensemencer 10 ml de milieu LB + ampicilline 100 μg/ml avec une colonie de bactéries BL21 (λDE3) porteuse du plasmide d'intérêt.
- incuber sous agitation une nuit à 37°C
- ensemencer 300 ml de milieu LB + ampicilline 100 μg/ml avec 3 ml de préculture
- incuber sous agitation à 37°C
- mesurer régulièrement la densité optique de la culture à 600 nm. Lorsqu'elle se situe entre 0,4 et 0,6, induire la production de la protéine d'intérêt par ajout d'IPTG 0,5mM.
- incuber sous agitation à 37°C pendant 3 heures
- centrifuger la culture à 5000 rpm pendant 15 minutes
- récupérer le culot bactérien et resuspendre dans un tampon PBS (Phosphate Buffer Saline : NaCl 137 mM, KCl 2,7 mM, KH_2PO_4 1,5 mM et Na_2HPO_4 8 mM - ajuster à pH 7,9).
-disrupter les bactéries par dépressurisation ("Constant Cell Disruption System", IKS Labs Equipment).
- récupérer le lysat cellulaire et centrifuger 30 minutes à 15 000 rpm.

- récupérer la phase liquide (fraction soluble) et recentrifuger 20 minutes à 15 000 rpm.
- récupérer la phase liquide (fraction soluble)
- pooler les phases liquides récoltées lors des deux centrifugations successives (fraction soluble finale)
- resuspendre les culots dans du tampon PBS (fraction insoluble) (tampon PBS : Phosphate Buffer Saline : NaCl 137 mM, KCl 2,7 mM, KH_2PO_4 1,5 mM et Na_2HPO_4 8 mM - ajuster à pH 7,9).

4.3. Protocole de surexpression pour l'optimisation de la température de production

- ensemencer 5 ml de milieu LB + ampicilline 100 µg/ml avec une colonie de bactéries BL21 (λDE3) porteuse du plasmide d'intérêt.
- incuber sous agitation une nuit à 37°C
- ensemencer 300 ml de milieu LB + ampicilline 100 µg/ml avec 3 ml de préculture
- incuber sous agitation à 37°C
- mesurer régulièrement la densité optique de la culture à 600 nm. Lorsqu'elle se situe entre 0,4 et 0,7, induire la production de la protéine d'intérêt par ajout d'IPTG 0,5mM.
- incuber sous agitation soit à 37°C, soit à 23°c soit à 18°C
- pour chaque culture cultivée à une température préçise, prélever 20 ml de culture après 1h, 2h, 3h, 4h, 7h, 9h et 21h d'induction.
- pour chaque aliquot prélevé : centrifuger la culture à 5000 rpm pendant 15 minutes
- récupérer le culot bactérien et resuspendre dans 3 ml de tampon PBS
- disrupter les bactéries par dépressurisation ("Constant Cell Disruption System", IKS Labs Equipment).
- récupérer le lysat cellulaire et centrifuger 30 minutes à 15 000 rpm.
- récupérer la phase liquide (fraction soluble) et recentrifuger 20 minutes à 15 000 rpm.
- récupérer la phase liquide (fraction soluble)
- pooler les phases liquides venant des deux centrifugations (fraction soluble finale)
- resuspendre les culots dans 3 ml de tampon PBS (fraction insoluble).
- aliquoter les fractions soluble et insoluble par 3 ml pour les analyses en SDS-PAGE.

4.4. Protocole de surexpression pour l'optimisation de la concentration en IPTG

- ensemencer 10 ml de milieu LB + ampicilline 100 µg/ml avec une colonie de bactéries BL21 (λDE3) porteuse du plasmide d'intérêt.
- incuber sous agitation une nuit à 37°C
- ensemencer 300 ml de milieu LB + ampicilline 100 µg/ml avec 3 ml de préculture
- incuber sous agitation à 37°C
- mesurer régulièrement la densité optique de la culture à 600 nm. Lorsqu'elle se situe entre 0,4 et 0,7, induire la production de la protéine d'intérêt par ajout d'IPTG 1 mM, 0,5mM, 0,1 mM ou 0,05 mM.
- incuber sous agitation à 18°C
- pour chaque culture, prélever 20 ml de culture après 7h, 9h et 21h d'induction.
- pour chaque aliquot prélevé : centrifuger la culture à 5000 rpm pendant 15 minutes
- récupérer le culot bactérien et resuspendre dans 3 ml de tampon PBS
- disrupter les bactéries par dépressurisation ("Constant Cell Disruption System", IKS Labs Equipment).
- récupérer le lysat cellulaire et centrifuger 30 minutes à 15 000 rpm.
- récupérer la phase liquide (fraction soluble) et recentrifuger 20 minutes à 15 000 rpm.
- pooler les phases liquides provenant des deux centrifugation (fraction soluble)
- resuspendre les culots dans 3 ml de tampon PBS (fraction insoluble).
- aliquoter les fractions soluble et insoluble par 1 ml pour les analyses en SDS-PAGE.

4.5. Protocole de surexpression pour l'optimisation de la composition du milieu de culture

Milieux de culture testés :

LB + NaCl 0,3M, soit NaCl 0,45M final
LB + sucrose 0,6M
LB + sorbitol 0,44M + glycine bétaïne 2,5 mM
LB + éthanol 4%
Milieu très riche (2% bactopeptone, 0,2% Na_2HPO_4, 0,1% KH_2PO_4, 0,8% NaCl, 1,5% yeast extract, 0,2% glucose)

- ensemencer 5 x 10 ml de milieu LB + ampicilline 100 µg/ml avec cinq colonies de bactéries BL21 (λDE3) porteuse du plasmide d'intérêt.
- incuber sous agitation à une nuit à 37°C

- ensemencer 300 ml de chaque milieu additionné d'ampicilline 100 µg/ml avec 3 ml de préculture
- incuber sous agitation à 37°C
- mesurer régulièrement la densité optique de la culture à 600 nm. Lorsqu'elle se situe entre 0,4 et 0,7, induire la production de la protéine d'intérêt par ajout d'IPTG 1 mM.
- incuber sous agitation à 18°C
- pour chaque culture, prélever 20 ml de culture après 7h, 9h et 21h d'induction.
- pour chaque aliquot prélevé : centrifuger la culture à 5000 rpm pendant 15 minutes
- récupérer le culot bactérien et resuspendre dans 3 ml de tampon PBS
-disrupter les bactéries par dépressurisation ("Constant Cell Disruption System", IKS Labs Equipment).
- rincer le disrupteur avec 2 ml de tampon PBS et les ajouter au lysat cellulaire
- centrifuger le lysat cellulaire 30 minutes à 15 000 rpm.
- récupérer la phase liquide (fraction soluble) et recentrifuger 20 minutes à 15 000 rpm.
- pooler les phases liquides provenant des deux centrifugation (fraction soluble)
- resuspendre les culots dans 5 ml de tampon PBS (fraction insoluble).
- aliquoter les fractions soluble et insolble par 1 ml pour les analyses en SDS-PAGE.

4.6. Protocole de surexpression de la thiorédoxine-L234 et des protéines mutantes thiorédoxine-EDE-L234, thiorédoxine-NG-L234 et thiorédoxine-V1V2V3-L234

- ensemencer 10 ml de milieu LB + ampicilline 100 µg/ml avec une colonie de bactéries BL21 (λDE3) porteuse du plasmide d'intérêt.
- incuber sous agitation à une nuit à 37°C
- ensemencer 300 ml de milieu LB-NaC 0,3M + ampicilline 100 µg/ml avec 3 ml de préculture
- incuber sous agitation à 37°C
- mesurer régulièrement la densité optique de la culture à 600 nm. Lorsqu'elle se situe entre 0,4 et 0,7, induire la production de la protéine d'intérêt par ajout d'IPTG 1 mM.
- incuber sous agitation à 18°C pendant 21h
- récupérer le culot bactérien et resuspendre dans 5 ml de tampon PBS
- disrupter les bactéries par dépressurisation ("Constant Cell Disruption System", IKS Labs Equipment).
- récupérer le lysat cellulaire et centrifuger 30 minutes à 15 000 rpm.
- récupérer la phase liquide (fraction soluble) et recentrifuger 20 minutes à

15 000 rpm.
- pooler les phases liquides provenant des deux centrifugation (fraction soluble)
- resuspendre les culots dans 5 ml de tampon PBS (fraction insoluble).

5. Purification sur colonne de chélation : chromatographie d'affinité par immobilisation de métaux (IMAC)

- couler une colonne de 2,5 ml de gel résine "His.Bind" (Novagen). Il s'agit d'une matrice d'agarose sur laquelle est fixée de façon covalente l'acide iminodiacétique (IDA) qui permet la fixation des ions métalliques (ici Ni^{++})
- laver la colonne avec 3 volumes d'eau bidistillée
- laver la colonne avec 4 volumes de "charge buffer" (50 mM $NiSO_4$)
- laver la colonne avec 3 volumes de "binding buffer" (5 mM imidazole, 0,5M NaCl, 20 mM Tris-Hcl pH 7,9)
- déposer l'échantillon à purifier
- laver la colonne avec 10 volumes de "binding buffer" (5 mM imidazole, 0,5M NaCl, 20 mM Tris-Hcl pH 7,9)
- laver la colonne avec 6 volumes de "wash buffer" (60 mM imidazole, 0,5M NaCl, 20 mM Tris-Hcl pH 7,9)
- éluer la protéine d'intérêt avec 6 volumes de tampon Tris 20 mM pH 7,9 EDTA 50 mM
- dialyser la protéine d'intérêt dans du tampon PBS 12h à 4°C dans 1000 fois le volume d'élution

6. Dosage protéique

Tous les dosages protéiques de ce travail ont été réalisés par la méthode de micro-BCA développée par la firme Pierce Chemical Co selon la méthode publiée par Smith (Smith et al., 1985).
L'acide bicinchonique (BCA) réagit avec les complexes de Cu2+ et de protéines qui peut être quantifiée à 562 nm. En formant de tels complexes, il prend une couleur pourpre typique. C'est une méthode sensible et rapide qui résiste aux détergents comme le Triton ou le SDS.
Les mesures ont été réalisées en triplicat et à deux concentrations différentes.

7. Précipitation des protéines au sulfate d'ammonium

- préparer une solution saturée en sulfate d'ammonium à 4°C (750 g par litre d'eau désioninsée)
- aliquoter 100 à 200 µg de protéine d'intérêt dans des eppendorfs. Ajouter dans chaque eppendorf du sulfate d'ammonium saturé de façon à obtenir une

concentration finale en sulfate d'ammonium de 0 %, 20 %, 25 %, 30 %, 40 %, 45 %, 50 % ou 55 % pour un volume final de 1,5 ml.
- incuber sous agitation une nuit à 4°C
- centrifuger 15 minutes à 14 000 rpm
- prélever 1 ml de surnageant et lire l'absorbance à 280 nm.

8. Estimation de l'affinité des protéines d'intérêt pour le DEAE-Sépharose

Toutes les expériences de purification ont été réalisée sur une FPLC (Fast Protein Liquid Chromatography) "AKTA Purifier" (Amersham Pharmacia Biotech).
Une colonne de 2 ml de gel DEAE-Sépharose Fast Flow (Amersham Pharmacia Biotech) a été utilisée pour les expériences. Certaines caractéristiques techniques du gel DEAE-Sépharose Fast Flow sont reprises au tableau 32.

Capacité ionique totale	0,11 - 0,16 mmol/ml gel	
Capacité de liaison	thyroglobuline (poids moléculaire : 669000 Da)	3,1 ml/ml
	HSA (poids moléculaire : 68000 Da)	110 mg/ml
	a-lactalbumine (poids moléculaire : 14 300 Da)	100 mg/ml
Structure des billes	agarose 6%	
Taille des billes	45 - 165 µm	
Taille moyenne des billes	90 µm	
Gamme de pH tolérée	2 - 9	

Tableau 32 Caractéristiques techniques du gel DEAE-Sépharose Fast Flow
Pour toutes les expérience, le débit est fixé à 2 ml/min et la pression à 0,5 Bar
Les fractions sont automatiquement récoltées par volume de 5 ml.

8.1. Estimation de l'affinité des protéines de fusion thiorédoxine-fragments de domaine lorsqu'elles sont pré-purifiées ou présentes dans la fraction soluble d'*E. coli*

- Equilibrer la colonne avec 20 ml de tampon phosphate 50 mM pH 8
- injecter automatiquement l'échantillon dans un volume de 3 ml
- laver la colonne avec 12 ml de tampon phosphate 50 mM pH 8
- appliquer un gradient linéaire de NaCl 0-2 M en 40 ml
- laver la colonne avec 15 ml de tampon phosphate 50 mM pH 8 NaCl 2M
- laver la colonne avec 15 ml de tampon phosphate 50 mM pH 8
- laver la colonne avec 15 ml de tampon phosphate 50 mM pH 8 choline 2%.

8.2 Mise en évidence des différences d'affinité pour le DEAE-Sépharose des protéines purifiées thiorédoxine-L234, thiorédoxine-EDE-L234, thiorédoxine-NG-L234 et thiorédoxine-V1V2V3-L234

- Equilibrer la colonne avec 20 ml de tampon phosphate 50 mM pH 8
- injecter automatiquement l'échantillon dans un volume de 3 ml
- laver la colonne avec 16 ml de tampon phosphate 50 mM pH 8
- appliquer un gradient linéaire de choline 0 - 800 mM en 40 ml
- laver la colonne avec 18 ml de tampon phosphate 50 mM pH 8 choline 800 mM

9. Dosage des bandes protéiques en SDS-PAGE par "scanning" densitométrie après coloration du gel au Bleu de Coomasie

Une électrophorèse en conditions dénaturantes est réalisée pour les différents échantillons protéiques à analyser.
- Déposer les fractions sur gel pré-coulé Criterion XT Bis-Tris 12% (Bio-Rad)
- Réaliser l'électrophorèse dans un tampon MOPS (MOPS 50 mM, Tris Hcl 50 mM pH 7,7, SDS 3,5 mM, EDTA 1mM) à 400 mA et 200V.
- Laver le gel 3 x 10 minutes dans 200 ml d'eau bidistillée.
- colorer le gel pendant 1 heure à l'aide d'une solution de bleu de Coomasie "Bio-Safe Coomassie Stain" (Bio-Rad)
- décolorer le gel quelques heures dans de l'eau bidistillée.
-scanner le gel sur un scanner
- réaliser une analyse densitométrique des bandes d'intérêt .

10. Analyses en Western blot des fractions protéiques

Les fractions protéiques récoltées lors des différents essais de purification ont été soumises à une électrophorèse dans un gel SDS-PAGE 12% en conditions dénaturantes puis transférées sur une membrane de nitrocellulose Hybond-C (Amersham Pharmacia Biotech) par un transfert semi-sec.
L'immunodétection des protéines de fusion a été réalisée avec l'anticorps 3G12 (1/1000) dirigé contre le *tag* 6 Histidines. La liaison de l'anticorps primaire a été détectée par chémoluminescence à l'aide d'anticorps secondaire conjugué à la peroxydase (1/5000) et des réactifs de Western blotting ECL RPN2209 (Amersham Pharmacia Biotech).

11. Mesure de la fluorescence des protéines de fusion thiorédoxine-fragment du domaine de liaison ClytA

Toutes les mesures de fluorescence ont été réalisées sur un spectro-fluorimètre "SLM-aminco Bowman series 2" et les données traitées par le logiciel AB2 (Life Sciences International (Europe) Ltd.).

- Dans une série d'eppendorfs, aliquoter l'équivalent de 200 µg de la protéine d'intérêt resuspendue dans du tampon PBS pH 7,9.
- ajouter du chlorure de guanidium 1M, 2M, 3M, 4M, 5M, 6M, 7M ou 8M et incuber 30 minutes à température ambiante.
- pour chaque échantillon, mesurer la fluorescence en fixant la longueur d'onde d'exicitation à 282 nm et la longueur d'onde d'émission à 340 nm.

12. Programmes bioinformatiques

Dans cette partie, nous rappelons brièvement le but et le principe général des différents programmes bioinformatiques utilisés.

12.1. L'algorithme PSI-Blast

L'algorithme PSI-Blast est dédicacé à la recherche de séquences similaires à une séquence d'intérêt dans des banques de données (Altschul et al., 1997).
De manière générale, ce programme fonctionne en cinq étapes :
1) Lors d'une première itération, l'algorithme Blast cherche des séquences présentant des similarités locales avec la séquence d'intérêt.
2) Puis, chaque séquence sélectionnée est alignée à la séquence d'intérêt dans un alignement multiple
3) cet alignement sert de base à la création d'un profil et d'une matrice de scores spécifiques de la position
4) une recherche de nouvelles séquences similaires dans les banques de données est réalisée en utilisant la matrice de scores spécifique de la position
5) si de nouvelles séquences sont sélectionnées, le programme retourne au deuxième point ; sinon il arrête.

12.2. L'algorithme SAPS

Sur base de critères statistiques, cet algorithme analyse notamment la composition relative de la séquence d'intérêt en chaque acide aminé, la distribution des résidus chargés et d'autres types de résidus, la présence de structures répétées, l'existence de périodicités locales dans la séquence aidant à mettre en évidence des structures régulières potentielles et l'espacement entre divers types d'acides aminés (Brendel et al., 1992).

12.3. Programme d'alignement multiple ClustalW

Ce programme réalise des alignements multiples de séquences (Thompson et al., 1994).
Face à un set de séquences, le programme ClustalW aligne 2 à 2 toutes les séquences, les unes contre les autres et attribue un score à chaque alignement réalisé. Puis, Clustal W sélectionne l'alignement pairé présentant le plus haut score (soit les séquences A et B), le fixe et en définit un profil. L'étape suivante consiste à aligner toutes les séquences restantes et le profil créé les uns contre les autres. Si le plus haut score correspond à l'alignement de A-B avec une séquence C, cet alignement est sélectionné et un nouveau profil est

défini. Par contre, si le plus haut score d'alignement concerne les séquences C et D en excluant l'alignement AB, ClustalW va aligné C-D, et en faire un profil. L'étape d'après consiste à tester les alignements entre toutes les séquences restantes et en faisant participer les deux profils.
L'arbre guide de ClustalW est une mémoire de l'ordre dans lequel les séquences ont été alignées et constitue donc un certain reflet de leur degré de similarité.

12.4. Match-Box

Match-Box est un programme d'alignement de séquences, basé sur des critères strictement statistiques (Depiereux and Feytmans, 1992). Il contourne la nécessité d'introduire des pénalités de "*gap*" puisque, dans la méthode, les *gaps* sont le résultat de l'alignement et non un paramètre gouvernant la procédure d'alignement. Il fournit un score de confiance pour chaque position alignée.
Ce programme d'alignement multiple local fonctionne en deux étapes (Lambert, 2003).
La première, appelée EXPLORE, effectue une analyse statistique sur les séquences pour déterminer leur similarité globale. Cette étape permet de discriminer entre deux séquences présentant une réelle similarité et deux séquences alignées par hasard.
La deuxième partie, appelée ALIGN, réalise l'alignement des séquences. Elle peut à son tour être divisée en trois étapes. :
le SCANNING consistant de nouveau en une analyse des séquences et calculant les meilleures valeurs de paramètres utilisés dans les étapes qui suivent ;
le MATCHING qui recherche des segments de longueur fixe conservés dans toutes les séquences et génère des boîtes.
Le SCREENING sélectionne les boîtes les plus appropriées pour la réalisation de l'alignement final.
Quatre itérations matching/screening sont réalisées avant qu'un alignement final ne soit proposé, avec un indice de confiance pour chaque position alignée.

12.5. Dialign 2

Par rapport aux méthodes standards d'alignement qui reposent sur la comparaison de résidus individuels et sur l'ajout de pénalités en cas d'introduction d'un "*gap*", DIALIGN construit des alignements pairés et multiples en comparant des segments complets de séquences de même longueur. Ces paires de segments sont appelées des diagonales (Morgenstern, 1999).
Dans une première étape, il y a formation d'un alignement entre chaque paire de séquences retenues pour l'alignement multiple. Les diagonales de similarités sont classifiées selon leur score et leur degré de superposition avec d'autres diagonales.
La construction de l'alignement multiple s'effectue à partir des diagonales. La diagonale ayant le plus haut score est utilisée comme séquence de départ. Ensuite, si la diagonale suivante de la classification est consistante avec l'alignement, elle est rajoutée à l'alignement primaire. Cette opération est effectuée pour toutes les diagonales.
Ces procédures sont effectuées jusqu'à ce que aucune nouvelle diagonale ne puisse être trouvée. L'étape finale consiste à insérer des gaps dans les séquences afin que tous les résidus soient alignés correctement.
Les alignements sont donc considérés comme des collections de diagonales qui doivent être cohérentes entre-elles Si deux séquences doivent être alignées on dit qu'il y a cohérence si les paires de segments sont ordonnées de telle façon que, pour n'importe quelle paire de diagonales, les extrémités d'une diagonale n'empiètent pas sur les extrémités d'une autre diagonale.
Pour trouver les bonnes collections de diagonales, on attribue un score à toutes les diagonales. Il faut ensuite trouver la collection de diagonales qui possède le score global le plus élevé.

12.6. Blockmaker

Ce serveur sert à identifier et fournir des régions conservées communes à une famille de protéines. Il construit donc des blocs de similarités dans des sets de séquences (Henikoff et al., 1995). Pour ce faire, il utilise deux algorithme de recherche de motifs déjà existants: MOTIF et GIBBS."motif finder basé sur spaced triplets et une adaptation automatique du motif finder basé sur l'échantillonnage de Gibbs.
Chaque algorithme possède son propre système de scores et fournit des sets de résultats indépendants que l'on peut comparer. Les blocs découverts par les deux méthodes peuvent être considérés comme valables tandis que ceux trouvés par un seul algorithme doivent être confirmés .
Ces deux algorithmes assurent que toutes les séquences d'un groupe partagent au moins un motif et donc, les programmes continuent jusqu'à ce qu'ils en trouvent au moins un , même dans des séquences générées aléatoirement.

12.7. PSI-PRED

Ce programme est utilisé pour prédire des structures secondaires (Jones, 1999).
Cette méthode utilise des profils intermédiaires créés par le programme d'alignement PSI-Blast pour générer, après trois itérations, une matrice de scores spécifiques. Les données de fréquence de chaque résidu à chaque position, contenues dans la matrice sont ensuite introduites dans un réseau neuronal dont le but est alors prédire les structures secondaires de la séquence d'intérêt.
Donc, dans une première étape, PSI-BLAST réalise trois itérations pour aller rechercher des séquences similaires dans les banques de données puis crée un profil sur base du dernier alignement multiple. Ce profil représente le pourcentage de chaque résidu à chaque position de l'alignement.
Dans un deuxième temps, la matrice du profil est soumise à un réseau neuronal après entraînement de celui-ci avec des cas-tests. On fournit au réseau une fenêtre de N résidus entourant l'acide aminé pour lequel on veut prédire la structure secondaire. Si N est de 6, la fenêtre est de 13. Chaque résidu est représenté par un vecteur de 21 valeurs : la fréquence des 20 acides aminés à cette position et une valeur pour les bords (0 ou 1) quand la fenêtre va de 7 à 12 résidus.
Le système contient donc au total 21 x 13 neurones d'entrée. Dans le système, il y a une couche cachée puis la couche de sortie à trois états : hélices, brins ou *loops*.
Le résultat de la prédiction est théoriquement la probabilité de chaque état (hélice, brin ou *loop*) pour chaque résidu.
Pour chaque résidu, le programme analyse ensuite les valeurs des trois états pour décider lequel il faut sélectionner. Des filtres sont également appliqués.
Exemples de filtres : impossible d'avoir une hélice avec moins de 4 résidus, impossible d'avoir un brin avec moins de 2 résidus, correction d'une prédiction de style HHHHHHBBHHHH en HHHHHHHHHHHH

12.8. Les serveurs de pseudo-*threading*

Les serveurs 3D-PSSM (Kelley et al., 1999) et UCLA.DOE (Fischer, 1999) se basent sur le principe suivant. D'une part, chaque structure de la banque de donnée est définie comme une succession de propriétés (ou de profils) associées à chaque résidu de cette structure. Les profils définis servent ensuite à créer des matrices de scores spécifiques. Les propriétés analysées sont, notamment, le type de résidus à chaque position, la structure secondaire et l'accessibilité au solvant du résidu. En parallèle, les serveurs créent aussi des matrices de scores spécifiques pour la séquence d'intérêt par alignements multiples avec des séquences homologues retrouvées dans les banques de données par le programme PSI-Blast. Outre le type de résidu à chaque position, les matrices de scores attribuées à la séquence d'intérêt illustrent entre autre la préférence de structure secondaire et l'accessibilité au solvant. La recherche de structures potentielles pour une séquence donnée se fait par comparaison des profils de la séquence aux profils calculés pour les structures de la banque.

12.9. Les serveurs de *threading*

Ils se basent sur des considérations énergétiques (utilisation de potentiels statistiques) . La séquence d'intérêt est alignée à chacune des structures d'une librairie de *folds* et le programme recherche l'alignement séquence-structure qui est le plus énergétiquement favorable.

ANNEXE 1

ETABLISSEMENT D'UN CONSENSUS DE SÉQUENCE DES MOTIFS REPETES CONSTITUANT LES DOMAINES DE LIAISON À LA CHOLINE

ANNEXE 1

	CLUSTALW	DIALIGN2	BLOCKMAKER (MOTIF)	BLOCKMAKER (GIBBS)	Match-Box
A : position N° 1	**HYDROPHOBES 72,3%** dont 29,4% L et 27,8% V I>A	**HYDROPHOBES 63,5%** dont 27,8% L 22,2% V I>A	**HYDROPHOBES 63,2%** dont 28,8% L 22,4% V I>A	**HYDROPHOBES 62,4%** dont 28% L 22,4% V I>A	**HYDROPHOBES 61,9%** dont 27,8% L 22,2% V I>A
B : position N° 2	CHARGES 42,1% dont 38,1% K R>H>E POLAIRES 44,4% dont 34,1% Q N>S>T **Total ch/pol = 86,5**	CHARGES 42,9% dont 36,5% K H>R>E POLAIRES 48,4 % dont 32,5% Q N>S>T **Total ch/pol = 91,3** GAP GAP	CHARGES 44,8% dont 38,4% K H>R>E POLAIRES 44,6 % dont 30,4% Q 10,4% N>S>T **Total ch/pol = 89,4**	CHARGES 43,2% dont 36,8% K H>R>E POLAIRES 46,4 % dont 31,2% Q N>S>T **Total ch/pol = 89,6**	CHARGES 42,1% dont 35,7% K H>R>E POLAIRES 47,6 % dont 31,7% Q N>S>T **Total ch/pol = 89,7**
C : position N° 3	HYDROPHOBES 23% V>I>L>A AROMATIQUES 15,1% CHARGES 44,4% dont 34,9% D E>R>H POLAIRES 17,5% dont 11,9% N **total ch/pol = 61,9** **total phobes/arom = 38,1**	HYDROPHOBES 23% V>I>L>A AROMATIQUES 15,9% Y CHARGES 40,5% dont 29,4% D E>H>R POLAIRES 20,7% dont 12,7% N T>S>Q **total ch/pol = 61,2** **total phobes/arom = 38,9**	HYDROPHOBES 22,4% V>I>L>A AROMATIQUES 4,8 % CHARGES 40,8% dont 38,4% K POLAIRES 20,8% dont 12,8% N T>S=Q **total ch/pol = 61,6** **total phobes/arom = 27,2**	HYDROPHOBES 24% V>I>L=A AROMATIQUES 16% Y CHARGES 39,2% dont 29,6% D E>H>R POLAIRES 18,4 % dont 12,8% N T>S **total ch/pol = 57,6** **total phobes/arom = 40**	HYDROPHOBES 23,8% V>I>L=A AROMATIQUES 15,9% Y CHARGES 40,5% dont 30,2% D E>H>R POLAIRES 19,9% dont 11,9% N T>S=Q **total ch/pol = 60,4** **total phobes/arom = 39,4**

Tableau 33 Estimation de l'importance relative de chaque classe d'acides aminés à chaque position des cinq alignements multiples (ClustalW, Dialign2, Blockmaker (MOTIF), Blockmaker (GIBBS) et Match-Box).
Ligne A : exemple de classe systématiquement majoritaire pour les alignements puisque au moins 60% des résidus à cette position appartiennent à la même classe d'acides aminés (ici, les hydrophobes).
Ligne B : exemple de deux classes compatibles majoritaires puisque au moins 70% des résidus appartiennent à ces deux classes (ici, chargés ou polaires) et ce, pour tous les alignements.
Ligne C : exemple de position variable X où même en regroupant les classes d'acides aminés compatibles, le total des résidus appartenant à ces classes n'atteint pas 70%.
Les lettres représentent les symboles des 20 acides aminés
Ch - pol : acides aminés chargés - acides aminés polaires
Phobes - arom : acides aminés hydrophobes - acides aminés aromatiques

Classe	Acides aminés	Fréquence relative attendue de l'acide aminé	Fréquence relative attendue de la classe d'acides aminés
Résidus hydrophobes	alanine	0,09774	**0,27727**
	isoleucine	0,04414	
	leucine	0,06157	
	méthionine	0,01122	
	valine	0,0626	
Résidus aromatiques	phénylalanine	0,02392	**0,07943**
	tryptophane	0,0093	
	tyrosine	0,04621	
Résidus chargés positivement	lysine	0,11708	**0,15695**
	arginine	0,02835	
	histidine	0,01152	
Résidus chargés négativement	acide aspartique	0,08504	**0,18219**
	acide glutamique	0,09715	
Résidus polaires	sérine	0,06334	**0,21217**
	thréonine	0,06452	
	asparagine	0,05153	
	glutamine	0,03278	
Glycine	glycine	0,05404	**0,05404**
Cystéine	cystéine	0,00192	**0,00192**
Proline	proline	0,03602	**0,03602**

Tableau 34 Calcul de la fréquence relative d'utilisation des classes d'acides aminés à partir de la fréquence d'utilisation des acides aminés présents dans les domaines catalytiques des protéines se liant à la choline.

Position 6 de l'alignement ClustalW

classe d'acides aminés	fréquence attendue	X_{ij}	abondance moyenne attendue	P
hydrophobes	0,27727	29	35	0,04076888
aromatiques	0,07943	19	10	0,00285209
acides	0,15695	4	20	5,46447E-06
basiques	0,18219	52	23	1,07771E-09
polaires	0,21217	22	27	0,053408825
glycines	0,05404	0	7	0,000911946
cystéines	0,00192	0	0	0,784936434
prolines	0,03602	0	4	0,009830395

Position 6 de l'alignement Blockmaker (GIBBS)

classe d'acides aminés	fréquence attendue	X_{ij}	abondance moyenne attendue	P
hydrophobes	0,27727	30	35	0,050571526
aromatiques	0,07943	20	10	0,001316573
acides	0,15695	5	20	2,48225E-05
basiques	0,18219	44	23	3,75058E-06
polaires	0,21217	23	27	0,065038421
glycines	0,05404	0	7	0,000911946
cystéines	0,00192	0	0	0,784936434
prolines	0,03602	0	4	0,009830395

Position 6 de l'alignement Match-Box

classe d'acides aminés	fréquence attendue	X_{ij}	abondance moyenne attendue	P
hydrophobes	0,27727	30	35	0,050571526
aromatiques	0,07943	20	10	0,001316573
acides	0,15695	5	20	2,48225E-05
basiques	0,18219	46	23	5,9727E-07
polaires	0,21217	25	27	0,082596009
glycines	0,05404	0	7	0,000911946
cystéines	0,00192	0	0	0,784936434
prolines	0,03602	0	4	0,009830395

Position 6 de l'alignement Dialign2

classe d'acides aminés	fréquence attendue	X_{ij}	abondance moyenne attendue	P
hydrophobes	0,27727	29	35	0,04076888
aromatiques	0,07943	20	10	0,001316573
acides	0,15695	5	20	2,48225E-05
basiques	0,18219	46	23	5,9727E-07
polaires	0,21217	26	27	0,086408914
glycines	0,05404	0	7	0,000911946
cystéines	0,00192	0	0	0,784936434
prolines	0,03602	0	4	0,009830395

Position 6 de l'alignement Blockmaker (MOTIF)

classe d'acides aminés	fréquence attendue	X_{ij}	abondance moyenne attendue	P
hydrophobes	0,27727	28	35	0,031446615
aromatiques	0,07943	20	10	0,001316573
acides	0,15695	5	20	2,48225E-05
basiques	0,18219	46	23	5,9727E-07
polaires	0,21217	26	27	0,086408914
glycines	0,05404	0	7	0,000911946
cystéines	0,00192	0	0	0,784936434
prolines	0,03602	0	4	0,009830395

Tableau 35 Calcul de la probabilité d'observer une classe d'acides aminés à une position précise de l'alignement des 126 *repeats* de domaines de liaison à la choline.
Soit i et la classe d'acides aminés et j la position dans le consensus (ici, j correspond à la position n°6).
Pour chaque classe de résidus, une fréquence relative FR_{ij} est calculée à partir de la composition en acides aminés des domaines catalytiques des protéines se liant à la choline (voir tableau 5).
En approximant la probabilité P_{ij} par la fréquence relative Fr_{ij}, X_{ij} peut être définie comme une variable binomiale : v.a. Bi (n_{ij},P_{ij}) où n_{ij} est le nombre de positions alignées (ici, toujours 126).
Une fois la fréquence relative attendue FR estimée, l'abondance moyenne attendue peut être calculée par la formule n_{ij} x FR_{ij}.
Enfin, au départ de tables, on peut calculer la probabilité P d'observer par hasard le même nombre de fois cette classe d'acides aminés à cette position.
Si la probabilité P est très petite, l'abondance de cette classe d'acides aminés à cette position est significativement élevée ou faible par rapport à l'abondance moyenne attendue.
Dans notre exemple, on observe des probabilités très faibles pour les acides aminés acides et basiques. Mais la probabilité pour les acides aminés basiques est très faible parce qu'on observe très peu de résidus de ce type par rapport à ce qui est attendu tandis que la probabilité des résidus acides est très faible parce qu'on observe un nombre anormalement élevé de ces résidus par rapport à leur fréquence relative attendue.
Dans le cadre de l'établissement du consensus, la faible probabilité d'observer des résidus acides en si grand nombre tend à montrer que la présence de charges négatives à cette position n'est pas dûe au hasard. Dans le consensus final, on notera donc plutôt un résidu acide à la place du X.

Programme d'alignement utilisé	**ClustalW**				**Dialign2**			
Position dans le consensus	**6**	**11**	**17**	**20**	**9**	**12**	**18**	**21**
Volume de la chaîne latérale compris entre 0 et 40 Ä^3	39,7	37,3	64,3	66,6	34,2	42,8	64,3	61,8
Volume de la chaîne latérale compris entre 41 et 70 Ä^3	19,8	26,9	27	14,3	23	26,9	27	13,5
Volume de la chaîne latérale compris entre 71 et 100Ä^3	24,6	29,4	7,2	7,2	26,2	17,5	7,2	5,6
Volume de la chaîne latérale supérieur à 100 Ä^3	15,9	0,8	1,6	11,9	16,7	0,8	1,6	11,9

Programme d'alignement utilisé	**Match-Box**				**BlockMaker (Gibbs)**			
Position dans le consensus	**6**	**9**	**15**	**18**	**6**	**9**	**15**	**18**
Volume de la chaîne latérale compris entre 0 et 40 Å3	35,8	42	65,1	68,9	35,2	43,2	65,6	69,6
Volume de la chaîne latérale compris entre 41 et 70 Ä^3	21,4	27,7	27	13,5	21,6	27,2	26,4	13,6
Volume de la chaîne latérale compris entre 71 et 100 Ä^3	26,2	28,6	4	5,6	24	28	6,4	5,6
Volume de la chaîne latérale supérieur à 100 Ä^3	16,7	1,6	1,6	11,9	16,8	1,6	1,6	11,2

Programme d'alignement utilisé	**BlockMaker (MOTIF)**			
Position dans le consensus	**6**	**9**	**15**	**18**
Volume de la chaîne latérale compris entre 0 et 40 Ä^3	34,4	42,4	64	69,6
Volume de la chaîne latérale compris entre 41 et 70 Ä^3	22,4	27,2	26,4	12,8
Volume de la chaîne latérale compris entre 71 et 100 Ä^3	16	28,8	8	4,8
Volume de la chaîne latérale supérieur à 100 Ä^3	16,8	1,6	1,6	11,2

Tableau 36 Calcul de la fréquence relative (exprimée en %) des 20 acides aminés à des positions variables des consensus obtenus par alignements de 126 repeats par cinq programmes d'alignements. Les acides aminés ont été préalablement regroupés en quatre classes distinctes, en fonction du volume de leur chaîne latérale.

module 1-1	CspA rep3	CbpA rep 9
	CspB rep 2	CbpA rep 10
	CspB rep 3	PcpA rep 12
	CspD rep3	PcpC rep 3
	CPL1 rep4	PcpC rep 5
	CPL9 rep4	PspA rep 9
	LytA rep4	PspA rep 10
	EJ-1 rep 4	PspC rep 10
	HBL-3 rep 4	PspC rep 11
	PAL rep 1	SpsA rep 3
		SpsA rep 4
module 1-2	CspA rep 1	LytA rep 6
	CspB rep 1	HBL3 rep 6
	CspC rep 1	LytC rep 6
	CspD rep 1	LytA rep1
	HBL-3 rep 1	EJ-1 rep 1
	CPL1 rep1	LytA rep 3
	CPL9 rep1	EJ-1 rep 3
	HBL3 rep 3	
module 1-3	Lyt B rep 1	Lyt B rep 11
	Lyt B rep 3	Lyt B rep 12
	Lyt B rep 4	Lyt B rep 13
	Lyt B rep 6	PcpC rep 1
	Lyt B rep 8	LytC rep 1
	Lyt B rep 9	LytC rep 2
	Lyt B rep 10	LytC rep 4
module 1-4	CPL1 rep2	LytB rep 2
	CPL9 rep2	LytB rep 5
	LytA rep2	LytB rep 7
	EJ-1 rep 2	LytC rep 3
	HBL-3 rep 2	LytC rep 5
module 1-5	CPL9 rep 6	
	EJ-1 rep 6	
	PAL rep 3	
module 1-6	CPL1 rep 3	
	CPL9 rep 3	
module 2	CspA rep 2	PcpA rep6
	CspD rep 2	PcpA rep7
	PcpA rep1	PcpA rep8
	PcpA rep2	PcpA rep9
	PcpA rep3	PcpA rep10
	PcpA rep4	PcpA rep11
	PcpA rep5	
module 3	CPL1 rep5	HBL-3 rep 5
	CPL9 rep5	PAL rep 2
	LytA rep5	PcpC rep 2
	EJ-1 rep 5	PcpC rep 4
module 4	CbpA rep 1	PspC rep1
	PbcA rep1	SpsA rep1
	PspA rep 1	
module 5	CspA rep 4	CspC rep3
	CspB rep 4	CspC rep4
	CspB rep 5	CspC rep5
	CspB rep 6	CspD rep 4
	CspC rep2	CspD rep 5
module 6	CbpA rep2	PbcA rep 3
	CbpA rep3	PbcA rep 4
	CbpA rep4	PspC rep 2
	CbpA rep5	PspC rep 3
	CbpA rep7	PspC rep 6
	PspA rep 2	PspC rep 7
	PspA rep 3	PspC rep 8
	CbpA rep 4	PspC rep 9
	PbcA rep 2	

Tableau 37 Répartition des 126 motifs répétés en 11 modules en fonction de leur classification phylogénique par le programme ClustalW (Thompson et al., 1994).
Chaque motif répété est défini par le nom de la protéine à laquelle il appartient, suivi de "rep", suivi de la position à laquelle ce motif se situe dans le domaine de liaison. Par exemple, CspA rep 3 représente le troisième motif du domaine de liaison à la choline de la protéine Csp.

ANNEXE 2

RECHERCHE DE STRUCTURES TRIDIMENTIONNELLES CONNUES, POTENTIELLEMENT PROCHES DE LA STRUCTURE DES DOMAINES DE LIAISON À LA CHOLINE

ANNEXE 2

Nombre de *repeats* soumis aux programmes de reconnaissance de *fold*	**Position de ces *repeats* dans le domaine de liaison**	**Protéine de liaison à la choline concernée**	**Nombre total de repeats dans le domaine de liaison**	**Organisme d'origine de la protéine**
10	domaine complet	CbpA	10	*Streptococcus pneumoniae*
4	domaine complet	CspA	4	*Clostridium beiherinckii*
6	domaine complet	EJL	6	*phage EJ-1*
6	domaine complet	LytA	6	*Streptococcus pneumoniae*
15	domaine complet	LytB	15	*Streptococcus pneumoniae*
13	domaine complet	PcpA	13	*Streptococcus pneumoniae*
10	domaine complet	PspA	10	*Streptococcus pneumoniae*
4	domaine complet	SpsA	4	*Streptococcus pneumoniae*
2	les deux premiers *repeats* du domaine de liaison	CbpA	10	*Streptococcus pneumoniae*
		CspA	4	*Clostridium beiherinckii*
		EJL	6	*phage EJ-1*
		LytA	6	*Streptococcus pneumoniae*
		LytB	15	*Streptococcus pneumoniae*
		PcpA	13	*Streptococcus pneumoniae*
		PspA	10	*Streptococcus pneumoniae*
		SpsA	4	*Streptococcus pneumoniae*
2	n°2 et n°3	CbpA	10	*Streptococcus pneumoniae*
	n°2 et n°3	CspA	4	*Clostridium beiherinckii*
	n°2 et n°3	EJL	6	*phage EJ-1*
	n°2 et n°3	LytA	6	*Streptococcus pneumoniae*
	n°7 et n°8	LytB	15	*Streptococcus pneumoniae*
	n°2 et n°3	SpsA	4	*Streptococcus pneumoniae*
4	les quatre premiers *repeats* du domaine de liaison	CbpA	10	*Streptococcus pneumoniae*
		CspA	4	*Clostridium beiherinckii*
		EJL	6	*phage EJ-1*
		LytA	6	*Streptococcus pneumoniae*
		LytB	15	*Streptococcus pneumoniae*
		PcpA	13	*Streptococcus pneumoniae*
		PspA	10	*Streptococcus pneumoniae*
		SpsA	4	*Streptococcus pneumoniae*
4	n°2, n°3, n°4, n°5	EJL	6	*phage EJ-1*
	n°2, n°3, n°4, n°5	LytA	6	*Streptococcus pneumoniae*
	n°7, n°8, n°9, n°10	LytB	15	*Streptococcus pneumoniae*

Tableau 38 Liste des domaines de liaison à la choline soumis aux programmes de reconnaissance de fold 3D-PSSM (Kelley et al., 1999) et UCLA.DOE (Fischer, 1999)

domaine de liaison soumis	programme de threading	structure proposée	score 3D-PSSM E-value	score UCLA	source	classification structurale
EJ-1	UCLA.DOE	1fbr : fibronectine		5,8 (1° hit)	*Homo sapiens*	2 : brins β 10 : **β ribbon** 100 : topologie fibronectine 10 : superfamille 1fbr
SpsA	UCLA.DOE	1cgt (580-684) cyclodextrine glu-cosyltransférase		6,8 (1° hit)	*Bacillus circulans*	2 : brins β 60 : **β sandwich** 40 : topologie immunoglobulin-like 110 : superfamille 1cgt dom4
	3D-PSSM	1pamB2 cyclodextrine glu-canotransférase	0,81 (1° hit)		*Bacillus sp*	2 : brins β 60 : **β sandwich** 120 : topologie jelly rolls 160 : superfamille 1cgt dom2
CspA	UCLA.DOE	1ppi (404-496) a amylase		7,3 (1° hit)	*Sus scrofa*	2 : brins β 60 : **β sandwich** 120 : topologie jelly rolls 160 : superfamille 1cgt dom2
	3D-PSSM	1 cid-2 Cd4 (domaines 3-4)	1,06 (1° hit)		*Rattus rattus*	2 : brins β 60 : **β sandwich** 40 : topologie immunoglobulin-like 580 : superfamille 1tvd chain A

Tableau 39 Structures proposées par les serveurs de reconnaissance de *fold* UCLA.DOE et 3D-PSSM pour trois domaines de liaison à la choline. Pour le programme 3D-PSSM, le score est une E-value , c'est-à-dire la fréquence d'apparition, par hasard, d'une séquence dans une banque de données avec un score supérieur ou égal au score indiqué. Plus cette valeur est petite, plus le résultat est significatif. Pour le programme UCLA.DOE, le score est une valeu normalisée centrée autour de zéro. Plus cette valeur est élevée, plus le score est significatif.

Hit retenu	Protéines pour lesquelles ce hit a été sélectionné	Score et position du hit	Programme utilisé	Alignement séquence structure	Présence du motif brin β-coude-brin β dans la structure sélectionnée
1 fbr1	CbpA (repeats 1 et 2)	E value : 0,684 (50% confiance) **1° hit**	3DPSSM	3 brins β prédits alignés à des brins de la structure	6 motifs brin β -coude-brin β présents sur la portion de la structure sélectionnée
1 fbr1	CspA (repeats 1 et 2)	E value : 0,169 (80% confiance) **1° hit**	3DPSSM	3 brins β prédits alignés à des brins de la structure	
1 fbr1	LytB (repeats 1 et 2)	9,2 **2° hit**	UCLA.doe		
1 fbr1	LytB (repeats 1 et 2)	E value : 0,505 **3° hit**	3DPSSM	tous les acides aminés des deux *repeats* ne sont pas inclus dans l'alignement	
1 fbr1	PspA (repeats 1 et 2)	E value : 0,169 (80 % confiance) **1° hit**	3DPSSM	3 brins β prédits alignés à des brins de la structure	
1 fbr1	SpsA (repeats 1 et 2)	E value : 0,277 (70% confiance) **1° hit**	3DPSSM	3 brins β prédits alignés à des brins de la structure	

Tableau 40 Résultats obtenus pour les deux premiers *repeats* des domaines de liaison à la choline soumis aux programmes de reconnaissance de *fold* 3D-PSSM et UCLA.DOE. E-value signifie expected value.

Hit retenu	Protéines pour lesquelles la structure est sélectionnée	Score et position du hit	Programme utilisé	Nombre de brins b prédits pour le domaine de liaison à la choline	Présence de motifs brin β-coude-brin β dans la structure sélectionnée
1fbr1	CbpA (repeats 2 et 3)	E value : 0,81 2° hit	3DPSSM	3 brins β au lieu de 4	6 motifs présents sur la portion de la structure sélectionnée
1fbr1	EJ1(repeats 2 et 3)	E value : 0,051(95% confiance) 1° hit	3DPSSM	3 brins β au lieu de 4	
1fbr1	EJ1(repeats 2 et 3)	12,1 1° hit	UCLA.doe	3 brins β au lieu de 4	
1fbr1	CspA (repeats 2 et 3)	E value : 0,695 2° hit	3DPSSM	4 brins β sur 4	
1fbr1	LytA (repeats 2 et 3)	E value : 0,149 (80% confiance) 1° hit	3DPSSM	4 brins β et 1 hélice α au lieu de 4 brins β	
1fbr1	LytA (repeats 2 et 3)	8,3 2° hit	UCLA.doe	3 brins β sur 4	
1fbr1	LytB (repeats 7 et 8)	E value : 1,16 2° hit	3DPSSM	3 brins β sur 4	
1fbr1	SpsA (repeats 2 et 3)	E value : 0,686 (50% confiance) 3° hit	3DPSSM	4 brins β et 1 hélice α au lieu de 4 brins β	

Tableau 41 Résultats obtenus pour deux *repeats* de six domaines de liaison à la choline soumis aux programmes de reconnaissance de fold 3D-PSSM et UCLA.DOE. E-value signifie expected value.

Structure sélectionnée	Protéines pour lesquelles la structure est sélectionnée	Score et position du hit	Programme utilisé	Nombre de brins prédits pour le domaine de liaison à la choline	Présence du motif brin β-coude-brin β dans la structure sélectionnée	Création d'un modèle	Présence d'une cage de résidus aromatiques
1 qa9B	LytB	0,0467 (95% confiance) 1° hit	3DPSSM	9 brins β et 1 hélice α au lieu de 8 brins β	4 motifs présents		
	EJ1	0,794 2° hit	3DPSSM	7 brins β et 1 hélice α au lieu de 8 brins β		modèle construit sur l'alignement séquence/structure EJ1 -1qa9B	1 cage de résidus aromatiques impliquant : W et Y du 2° *repeat* D et W du 3° *repeat*
	LytA	1,58 3° hit	3DPSSM	9 brins β et 1 hélice α au lieu de 8 brins β			
	PcpA	1,35 4° hit	3DPSSM	7 brins β et 1 hélice α au lieu de 8 brins β			
	SpsA	1,07 3° hit	3DPSSM	7 brins β au lieu de 8			
1hnf2	PspA	E value : 0,754 1° hit	3DPSSM	8 brins β sur 8	3 motifs présents sur la portion de la structure sélectionnée		
	SpsA	E value : 0,462 1° hit	3DPSSM	7 brins β au lieu de 8		modèle construit sur l'alignement séquence/structure SpsA -1hnf2	1 cage de résidus aromatiques impliquant W, F et T du 1° *repeat,* W du 3° *repeat* et F du 4° *repeat*
	CbpA	E value : 1,2 4° hit	3DPSSM	7 brins β au lieu de 8			
	LytA	E value : 2 5° hit	3DPSSM	8 brins β et 1 hélice α au lieu de 8 brins β			
	PcpA	E value : 0,808 3° hit	3DPSSM	7 brins β et 1 hélice α au lieu de 8 brins β		modèle construit sur l'alignement séquence/structure PcpA - 1hnf2	pas de cage de résidus aromatiques
1fbr	CbpA	E value : 0,742 2° hit	3DPSSM	7 brins β au lieu de 8	8 motifs présents sur toute la séquence de 1fbr impliquée dans l'alignement séquence/ structure		
	LytB	3,6 5° hit	UCLA.doe	pas de prédiction sur la séquence complète mais seulement sur la portion alignée c'est-à-dire 4 brins β			
	EJ1	8,2 1° hit	UCLA.doe	pas de prédiction sur la séquence complète mais seulement sur la portion alignée c'est-à-dire 3 brins β			
	LytA	5,9 2° hit	UCLA.doe	pas de prédiction sur la séquence complète mais seulement sur la portion alignée c'est-à-dire 4 brins β			
	SpsA	1,34 4° hit	3DPSSM	7 brins β au lieu de 8		modèle construit sur l'alignement séquence/structure SpsA/1fbr	pas de cage de résidus aromatiques

Tableau 42 Résultats de threading obtenus pour les quatre premiers *repeats* des domaines de liaison à la choline soumis aux programmes de reconnaissance de *fold* 3D-PSSM et UCLA.DOE. E-value signifie expected value.

Hit retenu	Protéines pour lesquelles le hit sort	Score et position du hit	Programme utilisé	Nombre de brins β prédits pour le domaine de liaison à la choline	Présence de motifs brin β-coude-brin β dans la structure sélectionnée	Création d'un modèle	Présence d'une cage de résidus aromatiques dans le modèle construit
1svb	CbpA	E value : 0,774 1° hit	3DPSSM	8 brins β sur 8	2 motifs présents sur la portion de la structure sélectionnée	modèle construit sur l'alignement séquence/ structure CbpA -1svb	pas de cage de résidus aromatiques
1cid2	CbpA	E value : 0,97 2° hit	3DPSSM	8 brins β sur 8	4 motifs présents sur la portion de la structure sélectionnée	modèle construit sur l'alignement séquence/ structure CbpA -1cid2	une cage de résidus aromatiques formée de : Y du 1° *repeat* N du 2° *repeat* W,W et Y du 3° *repeat*
	CspA	E value : 1,26 4° hit	3DPSSM	8 brins β sur 8			pas de modèle
	EJ1	E value : 1,55 5° hit	3DPSSM	8 brins β et 1 hélice α au lieu de 8			pas de modèle
	LytA	E value : 1,28 3° hit	3DPSSM	7 brins β et 1 hélice α au lieu de 8			pas de modèle
	LytB	E value : 2,74 8° hit	3DPSSM	8 brins β sur 8			pas de modèle
1sppB	CspA	E value : 0,917 2° hit	3DPSSM	8 brins β sur 8	1 motif présent sur la portion de la structure sélectionnée	modèle construit sur l'alignement séquence/ structure CspA -1sppB	pas de cage de résidus aromatiques
1fbr	EJ1	E value : 1,4 4° hit	3DPSSM	8 brins β et 1 hélice α au lieu de 8	10 motifs présents sur la portion de la structure sélectionnée		pas de modèle
	LytA	E value : 0,821 2° hit	3DPSSM	7 brins β et 1 hélice α au lieu de 8			pas de modèle
	SpsA	E value : 1,02 2° hit	3DPSSM	7 brins β au lieu de 8		modèle construit sur l'alignement séquence/ structure SpsA - 1fbr	pas de cage de résidus aromatiques

Tableau 43 Résultats obtenus pour quatre *repeats* des domaines de liaison à la choline soumis aux programmes de reconnaissance de *fold* 3D-PSSM et UCLA.DOE. E-value signifie expected value

ANNEXE 3

ANALYSE DU CONSENSUS DE *REPEATS* SUR BASE DE DEUX STRUCTURES DE DOMAINES DE LIAISON A LA CHOLINE

ANNEXE 3

Distance maximale entre deux résidus pour la formation des ponts salins : 4,5 Å
Distance maximale entre deux résidus pour la définition des contacts hydrophobes : 4,5 Å
Distance maximale entre deux résidus pour la formation des ponts disulfures : 2,5 A°

L'histidine est considérée comme un résidu basique
La méthionine est considérée comme un résidu hydrophobe

Tableau 44 Critères utilisés par le logiciel MOE pour définir les différents types de contact existant entre les résidus d'une protéine.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Consensus de *repeats*	**pol/X**	**gly/X**	**arom**	**phobe/X**	**ch/pol**	**X**	**ch/pol**	**G**	**X**	**W**	**Y**	**Y**	**phobe**	**ch/pol**	**petit AA**	**ch/pol**	**G**	**petit AA**	**phobe**	**phobe**
Séquence de 4 *repeats* du domaine de liaison à la choline de la protéine SpaA																				
repeat 3	T	G	W	K	K	V	A	N	K	W	Y	Y	L	E	K	S	G	A	M	A
repeat 4	T	G	W	K	K	V	S	N	K	W	Y	Y	L	E	N	S	G	A	M	A
repeat 5	T	G	W	K	K	V	S	N	K	W	Y	Y	L	E	N	S	G	A	M	A
repeat 6	T	G	W	K	K	V	S	N	K	W	Y	Y	L	E	N	S	G	A	M	A
Séquence de 4 *repeats* du domaine de liaison à la choline de la protéine PcpA																				
repeat 2	T	G	W	V	K	D	K	G	L	W	Y	Y	L	N	E	S	G	S	M	A
repeat 3	T	G	W	V	K	D	K	G	L	W	Y	Y	L	N	E	S	G	S	M	A
repeat 4	T	G	W	V	K	D	K	G	L	W	Y	Y	L	N	E	S	G	S	M	A
repeat 5	T	G	W	V	K	D	K	G	L	W	Y	Y	L	N	E	S	G	S	M	A
Séquence de 4 *repeats* du domaine de liaison à la choline de la protéine PspC																				
repeat 5	T	G	W	L	Q	Y	N	G	S	W	Y	Y	L	N	A	N	G	D	M	A
repeat 6	T	G	W	F	Q	Y	N	G	S	W	Y	Y	L	N	A	N	G	D	M	A
repeat 7	T	G	W	F	Q	Y	N	G	S	W	Y	Y	L	N	A	N	G	D	M	A
repeat 8	T	G	W	L	Q	Y	N	G	S	W	Y	Y	L	N	S	N	G	A	M	V

Tableau 45 Séquences de fragments de domaines de liaison utilisées pour analyser le consensus de *repeats*. Ces séquences ont été soumises au logiciel MOE afin de créer des modèles homologues aux structures des domaines de liaison à la choline des protéines LytA et CPL1.

ANNEXE 4

DESCRIPTION DES FRAGMENTS DU DOMAINE CLYTA FUSIONNES A LA THIOREDOXI

	REPEAT 1																					REPEAT 2																							
fragment CLytA	L	T	I	E	T	G	W	Q	K	N	D	T	G	Y	W	Y	V	H	S	D	G	S	Y	P	K	D	K	F	E	K	I	N	G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
fragment L23456																	V	H	S	D	G	S	Y	P	K	D	K	F	E	K	I	N	G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
fragment L1234	L	T	I	E	T	G	W	Q	K	N	D	T	G	Y	W	Y	V	H	S	D	G	S	Y	P	K	D	K	F	E	K	I	N	G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
fragment L2345																	V	H	S	D	G	S	Y	P	K	D	K	F	E	K	I	N	G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
fragment L123	L	T	I	E	T	G	W	Q	K	N	D	T	G	Y	W	Y	V	H	S	D	G	S	Y	P	K	D	K	F	E	K	I	N	G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
fragment L234																	V	H	S	D	G	S	Y	P	K	D	K	F	E	K	I	N	G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A
fragment L345																																											G	Y	M	L	A
fragment L456																																														
fragment L23																	V	H	S	D	G	S	Y	P	K	D	K	F	E	K	I	N	G	T	W	Y	Y	F	D	S	S	G	Y	M	L	A

	REPEAT 3																					REPEAT 4																							
fragment CLytA	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A	T	G	W	K	K	I	A	D	K	W	Y	Y	F	N	E	E	G	A	M	K
fragment L23456	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A	T	G	W	K	K	I	A	D	K	W	Y	Y	F	N	E	E	G	A	M	K
fragment L1234	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A	T	G	W	K	K	I	A	D	K	W	Y	Y	F	N	E	E	G	A	M	K
fragment L2345	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A	T	G	W	K	K	I	A	D	K	W	Y	Y	F	N	E	E	G	A	M	K
fragment L123	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A																							
fragment L234	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A	T	G	W	K	K	I	A	D	K	W	Y	Y	F	N	E	E	G	A	M	K
fragment L345	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A	T	G	W	K	K	I	A	D	K	W	Y	Y	F	N	E	E	G	A	M	K
fragment L456																		G	E	M	A	T	G	W	K	K	I	A	D	K	W	Y	Y	F	N	E	E	G	A	M	K
fragment L23	D	R	W	R	K	H	T	D	G	N	W	Y	W	F	D	N	S	G	E	M	A																							

Tableau 46 Séquence des fragments de domaine de liaison CLytA fusionnés à la thiorédoxine (suite page suivante).

	REPEAT 5	REPEAT 6
fragment CLytA	T G W V K Y K D T W Y Y L D A K E G A M V S	N A F I Q S A D G T G W Y Y L K P D G T L A
fragment L23456	T G W V K Y K D T W Y Y L D A K E G A M V S	N A F I Q S A D G T G W Y Y L K P D G T L A
fragment L1234	T G	
fragment L2345	T G W V K Y K D T W Y Y L D A K E G	
fragment L123		
fragment L234	T G	
fragment L345	T G W V K Y K D T W Y Y L D A K E G ***E L***	
fragment L456	T G W V K Y K D T W Y Y L D A K E G A M V S	N A F I Q S A D G T G W Y Y L K P D ***E L***
fragment L23		

	QUEUE C- TERMINALE
fragment CLytA	D R P E
fragment L23456	D R P E F T V E P D G
fragment L1234	
fragment L2345	
fragment L123	
fragment L234	
fragment L345	
fragment L456	
fragment L23	

Tableau 46 (suite) Séquence des fragments de domaine de liaison CLytA fusionnés à la thiorédoxine.
Les résidus indiqués en italique et en gras sont normalement absent du domaine de liaison CLytA et leur introduction résulte de la stratégie de clonage des fragments L345 et L456

BILIOGRAPHIE

BIBLIOGRAPHIE

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Research, 25, 3389-3402.

Baba, T. and Schneewind, O. (1996) Target cell specificity of a bacteriocin molecule: a C-terminal signal directs lysostaphin to the cell wall of *Staphylococcus aureus*. Embo J, 15, 4789-4797.

Balachandran, P., Brooks-Walter, A., Virolainen-Julkunen, A., Hollingshead, S.K. and Briles, D.E. (2002) Role of pneumococcal surface protein C in nasopharyngeal carriage and pneumonia and its ability to elicit protection against carriage of *Streptococcus pneumoniae*. Infect Immun, 70, 2526-2534.

Banci, L., Bertini, I., Ciulli, A., Fragai, L., Luchinat, C. and Terni, B. (2003) Expression and high yield production of the catalytic domain of matrix metalloprotease 12 and of an active mutant with increased solubility. Journal of Molecular Catalysis A : Chemical, 204-205, 401-408.

Behr, T., Fischer, W., Peter-Katalinic, J. and Egge, H. (1992) The structure of pneumococcal lipoteichoic acid. Improved preparation, chemical and mass spectrometric studies. Eur J Biochem, 207, 1063-1075.

Berry, A.M., Lock, R.A., Hansman, D. and Paton, J.C. (1989) Contribution of autolysin to virulence of *Streptococcus pneumoniae*. Infect Immun, 57, 2324-2330.

Berry, A.M. and Paton, J.C. (2000) Additive attenuation of virulence of *Streptococcus pneumoniae* by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins. Infect Immun, 68, 133-140.

Berry, A.M., Paton, J.C. and Hansman, D. (1992) Effect of insertional inactivation of the genes encoding pneumolysin and autolysin on the virulence of *Streptococcus pneumoniae* type 3. Microb Pathog, 12, 87-93.

Bowden, G.A. and Georgiou, G. (1990) Folding and aggregation of beta-lactamase in the periplasmic space of *Escherichia coli*. J Biol Chem, 265, 16760-16766.

Braun, L., Dramsi, S., Dehoux, P., Bierne, H., Lindahl, G. and Cossart, P. (1997) InlB: an invasion protein of *Listeria monocytogenes* with a novel type of surface association. Mol Microbiol, 25, 285-294.

Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, B.E. and Karlin, S. (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci U S A, 89, 2002-2006.

Briese, T. and Hakenbeck, R. (1985) Interaction of the pneumococcal amidase with lipoteichoic acid and choline. Eur J Biochem, 146, 417-427.

Briles, D.E., Tart, R.C., Swiatlo, E., Dillard, J.P., Smith, P., Benton, K.A., Ralph, B.A., Brooks-Walter, A., Crain, M.J., Hollingshead, S.K. and McDaniel, L.S. (1998) Pneumococcal diversity: considerations for new vaccine strategies with emphasis on pneumococcal surface protein A (PspA). Clin Microbiol Rev, 11, 645-657.

Brissette, J.L., Russel, M., Weiner, L. and Model, P. (1990) Phage shock protein, a stress protein of *Escherichia coli*. Proc Natl Acad Sci U S A, 87, 862-866.

Brooks-Walter, A., Briles, D.E. and Hollingshead, S.K. (1999) The pspC gene of *Streptococcus pneumoniae* encodes a polymorphic protein, PspC, which elicits cross-reactive antibodies to PspA and provides immunity to pneumococcal bacteremia. Infect Immun, 67, 6533-6542.

Cheng, Q., Finkel, D. and Hostetter, M.K. (2000) Novel purification scheme and functions for a C3-binding protein from *Streptococcus pneumoniae*. Biochemistry, 39, 5450-5457.

Comfort, D. and Clubb, R.T. (2004) A comparative genome analysis identifies distinct sorting pathways in gram-positive bacteria. Infect Immun, 72, 2710-2722.

Cossart, P. and Jonquieres, R. (2000) Sortase, a universal target for therapeutic agents against gram-positive bacteria? Proc Natl Acad Sci U S A, 97, 5013-5015.

Cundell, D.R., Gerard, N.P., Gerard, C., Idanpaan-Heikkila, I. and Tuomanen, E.I. (1995) *Streptococcus pneumoniae* anchor to activated human cells by the receptor for platelet-activating factor. Nature, 377, 435-438.

De Las Rivas, B., Garcia, J.L., Lopez, R. and Garcia, P. (2002) Purification and polar localization of pneumococcal LytB, a putative endo-beta-N-acetylglucosaminidase: the chain-dispersing murein hydrolase. J Bacteriol, 184, 4988-5000.

Depiereux, E. and Feytmans, E. (1992) MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. Comput Appl Biosci, 8, 501-509.

di Guan, C., Li, P., Riggs, P.D. and Inouye, H. (1988) Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein. Gene, 67, 21-30.

Di Guilmi, A.M. and Dessen, A. (2002) New approaches towards the identification of antibiotic and vaccine targets in *Streptococcus pneumoniae*. EMBO Rep, 3, 728-734.

Diaz, E., Lopez, R. and Garcia, J.L. (1990) Chimeric phage-bacterial enzymes: a clue to the modular evolution of genes. Proc Natl Acad Sci U S A, 87, 8125-8129.

Dopazo, J., Mendoza, A., Herrero, J., Caldara, F., Humbert, Y., Friedli, L., Guerrier, M., Grand-Schenk, E., Gandin, C., de Francesco, M., Polissi, A., Buell, G., Feger, G., Garcia, E., Peitsch, M. and Garcia-Bustos, J.F. (2001) Annotated draft genomic sequence from a *Streptococcus pneumoniae* type 19F clinical isolate. Microb Drug Resist, 7, 99-125.

Einhauer, A. and Jungbauer, A. (2001) The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. J Biochem Biophys Methods, 49, 455-465.

Fernandez-Tornero, C., Garcia, E., Lopez, R., Gimenez-Gallego, G. and Romero, A. (2002a) Two New Crystal Forms of the Choline-binding Domain of the Major Pneumococcal Autolysin : Insights into the Dynamics of the Active Homodimer. J. Mol. Biol., 321, 163-173.

Fernandez-Tornero, C., Lopez, R., Garcia, E., Gimenez-Gallego, G. and Romero, A. (2001) A novel solenoid fold in the cell wall anchoring domain of the pneumococcal virulence factor LytA. Nat Struct Biol, 8, 1020-1024.

Fernandez-Tornero, C., Ramon, A., Fernandez-Cabrera, C., Gimenez-Gallego, G. and Romero, A. (2002b) Expression, crystallization and preliminary X-ray diffraction studies on the complete choline-binding domain of the major pneumococcal autolysin. Acta Crystallogr D Biol Crystallogr, 58, 556-558.

Fischer, D. (1999) Modeling three-dimensional protein structures for amino acid sequences of the CASP3 experiment using sequence-derived predictions. Proteins, Suppl 3, 61-65.

Fischer, H. and Tomasz, A. (1985) Peptidoglycan cross-linking and teichoic acid attachment in *Streptococcus pneumoniae*. J Bacteriol, 163, 46-54.

Fischer, W. (2000) Phosphocholine of pneumococcal teichoic acids: role in bacterial physiology and pneumococcal infection. Res Microbiol, 151, 421-427.

Fischer, W., Behr, T., Hartmann, R., Peter-Katalinic, J. and Egge, H. (1993) Teichoic acid and lipoteichoic acid of *Streptococcus pneumoniae* possess identical chain structures. A reinvestigation of teichoid acid (C polysaccharide). Eur J Biochem, 215, 851-857.

Fischer, W., Markwitz, S. and Labischinski, H. (1997) Small-angle X-ray scattering analysis of pneumococcal lipoteichoic acid phase structure. Eur J Biochem, 244, 913-917.

Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M. and Sippl, M.J. (1995) Progress in fold recognition. Proteins, 23, 376-386.

Garcia, J.L., Diaz, E., Romero, A. and Garcia, P. (1994) Carboxy-terminal deletion analysis of the major pneumococcal autolysin. J Bacteriol, 176, 4066-4072.

Garcia, J.L., Garcia, E., Sanchez-Puelles, J.M. and Lopez, R. (1988) Identification of a lytic enzyme of *Clostridium acetobutylicum* that degrades choline-containing pneumococcal cell walls. FEMS Microbiol Lett, 52, 133-138.

Garcia, J.L., Sanchez-Beato, A.R., Medrano, F.J. and Lopez, R. (1998) Versatility of choline-binding domain. Microb Drug Resist, 4, 25-36.

Garcia, P., Garcia, E., Ronda, C., Lopez, R. and Tomasz, A. (1983) A phage-associated murein hydrolase in *Streptococcus pneumoniae* infected with bacteriophage Dp-1. J Gen Microbiol, 129, 489-497.

Garcia, P., Garcia, J.L., Garcia, E., Sanchez-Puelles, J.M. and Lopez, R. (1990) Modular organization of the lytic enzymes of *Streptococcus pneumoniae* and its bacteriophages. Gene, 86, 81-88.

Garcia, P., Gonzalez, M.P., Garcia, E., Lopez, R. and Garcia, J.L. (1999a) LytB, a novel pneumococcal murein hydrolase essential for cell separation. Mol Microbiol, 31, 1275-1277.

Garcia, P., Paz Gonzalez, M., Garcia, E., Garcia, J.L. and Lopez, R. (1999b) The molecular characterization of the first autolytic lysozyme of *Streptococcus pneumoniae* reveals evolutionary mobile domains. Mol Microbiol, 33, 128-138.

Georgiou, G. and Valax, P. (1996) Expression of correctly folded proteins in *Escherichia coli*. Curr Opin Biotechnol, 7, 190-197.

Giffard, P.M. and Jacques, N.A. (1994) Definition of a fundamental repeating unit in streptococcal glucosyltransferase glucan-binding regions and related sequences. J Dent Res, 73, 1133-1141.

Gilis, D. and Rooman, M. (2000) PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein Eng, 13, 849-856.

Gillespie, S.H., McWhinney, P.H., Patel, S., Raynes, J.G., McAdam, K.P., Whiley, R.A. and Hardie, J.M. (1993) Species of alpha-hemolytic streptococci possessing a C-polysaccharide phosphorylcholine-containing antigen. Infect Immun, 61, 3076-3077.

Gosink, K.K., Mann, E.R., Guglielmo, C., Tuomanen, E.I. and Masure, H.R. (2000) Role of novel choline binding proteins in virulence of *Streptococcus pneumoniae*. Infect Immun, 68, 5690-5695.

Graslund, S., Eklund, M., Falk, R., Uhlen, M., Nygren, P.A. and Stahl, S. (2002) A novel affinity gene fusion system allowing protein A-based recovery of non-immunoglobulin gene products. J Biotechnol, 99, 41-50.

Gunneriusson, E., Samuelson, P., Ringdahl, J., Gronlund, H., Nygren, P.A. and Stahl, S. (1999) Staphylococcal surface display of immunoglobulin A (IgA)- and IgE-specific in vitro-selected binding proteins (affibodies) based on *Staphylococcus aureus* protein A. Appl Environ Microbiol, 65, 4134-4140.

Hachem, B., Andrews, B.A. and Asenjo, J.A. (1996) Hydrophobic partitioning of proteins in aqueous two-phases systems. Enzyme and Microbial Technology, 19, 507-517.

Hammerschmidt, S., Tillig, M.P., Wolff, S., Vaerman, J.P. and Chhatwal, G.S. (2000) Species-specific binding of human secretory component to SpsA protein of *Streptococcus pneumoniae* via a hexapeptide motif. Mol Microbiol, 36, 726-736.

Harel, M., Schalk, I., Ehret-Sabatier, L., Bouet, F., Goeldner, M., Hirth, C., Axelsen, P.H., Silman, I. and Sussman, J.L. (1993) Quaternary ligand binding to aromatic residues in the active-site gorge of acetylcholinesterase. Proc Natl Acad Sci U S A, 90, 9031-9035.

Henikoff, S., Henikoff, J.G., Alford, W.J. and Pietrokovski, S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene, 163, GC17-26.

Hermoso, J.A., Monterroso, B., Albert, A., Galan, B., Ahrazem, O., Garcia, P., Martinez-Ripoll, M., Garcia, J.L. and Menendez, M. (2003) Structural basis

for selective recognition of pneumococcal cell wall by modular endolysin from phage Cp-1. Structure (Camb), 11, 1239-1249.

Holtje, J.V. and Tomasz, A. (1975) Specific recognition of choline residues in the cell wall teichoic acid by the N-acetylmuramyl-L-alanine amidase of *Pneumococcus*. J Biol Chem, 250, 6072-6076.

Hoskins, J., Alborn, W.E., Jr., Arnold, J., Blaszczak, L.C., Burgett, S., DeHoff, B.S., Estrem, S.T., Fritz, L., Fu, D.J., Fuller, W., Geringer, C., Gilmour, R., Glass, J.S., Khoja, H., Kraft, A.R., Lagace, R.E., LeBlanc, D.J., Lee, L.N., Lefkowitz, E.J., Lu, J., Matsushima, P., McAhren, S.M., McHenney, M., McLeaster, K., Mundy, C.W., Nicas, T.I., Norris, F.H., O'Gara, M., Peery, R.B., Robertson, G.T., Rockey, P., Sun, P.M., Winkler, M.E., Yang, Y., Young-Bellido, M., Zhao, G., Zook, C.A., Baltz, R.H., Jaskunas, S.R., Rosteck, P.R., Jr., Skatrud, P.L. and Glass, J.I. (2001) Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol, 183, 5709-5717.

Ingraham, J.L. and Ingraham, C.A. (eds.). (1995) Introduction to Microbiology. Wadsworth Publishing Company, Belmont, California.

Jarvik, J.W. and Telmer, C.A. (1998) Epitope tagging. Annu Rev Genet, 32, 601-618.

Jedrzejas, M.J. (2001) Pneumococcal virulence factors: structure and function. Microbiol Mol Biol Rev, 65, 187-207 ; first page, table of contents.

Jones, D.T. (1999) Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. J. Mol. Biol., 292, 195-202.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition. Nature, 358, 86-89.

Jonquières, R., Bierne, H., Fiedler, F., Gounon, P. and Cossart, P. (1999) Interaction between the protein InlB of *Listeria monocytogenes* and lipoteichoic acid : a novel mechanism of protein association at the surface of Gram-positive bacteria. Mol Microbiol, 34, 902-914.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22, 2577-2637.

Karlsson, E., Ryden, L. and Brewer, J. (1989) Ion Exchange Chromatography. In Janson, J.C. and Ryden, L. (eds.), Protein purification : principles, high resolution methods, and applications. VCH Publishers, Inc., New York.

Kelley, L.A., Maccallum, R. and Sternberg, M.J.E. (1999) Recognition of Remote Protein Homologies Using Three-Dimensional Information to Generate a Position Specific Scoring Matrix in the program 3D-PSSM. In Sorin Istrail, P.P., Michael Waterman (ed.), RECOMB 99, Proceedings of the Third Annual Conference on Computational Molecular Biology. The Association for Computing Machinery, NewYork, pp. 218-225.

Kilpper-Bätz, R., Wenzig, P. and Schleifer, K.H. (1985) Molecular relationship and classification of some viridans streptococci as *Stretococcus oralis* and amended description of *Streptococcus oralis* (Bridge and Sneath 1982). Int. J. Sist. Microbiol., 35, 482-498.

Kim, J.O. and Weiser, J.N. (1998) Association of intrastrain phase variation in quantity of capsular polysaccharide and teichoic acid with the virulence of *Streptococcus pneumoniae*. J Infect Dis, 177, 368-377.

Kirk, O., Borchert, T.V. and Fuglsang, C.C. (2002) Industrial enzyme applications. Curr Opin Biotechnol, 13, 345-351.

Kocks, C., Gouin, E., Tabouret, M., Berche, P., Ohayon, H. and Cossart, P. (1992) *L. monocytogenes*-induced actin assembly requires the actA gene product, a surface protein. Cell, 68, 521-531.

Koehl, P. and Delarue, M. (1994) Polar and nonpolar atomic environments in the protein core: implications for folding and binding. Proteins, 20, 264-278.

Krigbaum, W.R. and Komoriya, A. (1979) Local interactions as a structure determinant for protein molecules : II. Biochim. Biophys. Acta, 576, 204-228.

Kufer, T.A., Fritz, J.H. and Philpott, D.J. (2005) NACHT-LRR proteins (NLRs) in bacterial infection and immunity. Trends Microbiol, 13, 381-388.

Lambert, C. (2003) Développement d'une méthode automatique fiable de modélisation de la structure tridimensionnelle des protéines par homologie et application au protéome de Brucella melitensis. département de biologie. FUNDP, Namur.

Lambert, C., Wouters, J., De Bolle, X. and Depiereux, E. (2003) Biologie in silico : Point de Vue de Bioinformaticiens. Chimie nouvelle, 83, 106-111.

laVallie, E.R., DiBlasio, E.A., Kovacic, S., Grant, K.L., Schendel, P.F. and McCoy, J.M. (1993) A Thioredoxin Gene Fusion Expression System That Circumvents inclusion Body Formation in the *E. coli* Cytoplasm. Biotechnology, 11, 187-193.

Lee, G.Y., Zhu, J., Yu, L. and Yu, C.A. (1998) Reconstitution of cytochrome b-560 (QPs1) of bovine heart mitochondrial succinate-ubiquinone reductase. Biochim Biophys Acta, 1363, 35-46.

Lee, V.T. and Schneewind, O. (2001) Protein secretion and the pathogenesis of bacterial infections. Genes Dev, 15, 1725-1752.

Lemer, C.M., Rooman, M.J. and Wodak, S.J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. Proteins, 23, 337-355.

Lopez, R., Garcia, E., Garcia, P. and Garcia, J.L. (1995) Architecture and domain interchange of the pneumococcal cell wall lytic enzymes. Dev Biol Stand, 85, 273-281.

Lopez, R., Garcia, E., Garcia, P., Ronda, C. and Tomasz, A. (1982) Choline-containing bacteriophage receptors in *Streptococcus pneumoniae*. J Bacteriol, 151, 1581-1590.

Lu, Z., Murray, K.S., Van Cleave, V., LaVallie, E.R., Stahl, M.L. and McCoy, J.M. (1995) Expression of thioredoxin random peptide libraries on the *Escherichia coli* cell surface as functional fusions to flagellin: a system designed for exploring protein-protein interactions. Biotechnology (N Y), 13, 366-372.

Lunn, C.A., Kathjus, S., Wallace, B.J., kushner, S.R. and Pigiet, V. (1984) Amplification and purification of a plasmid-encoded thioredoxin from *Escherichia coli* K12. J. Biol. Chem., 259, 10469-10474.

Ma, D., Alberti, M., Lynch, C., Nikaido, H. and Hearst, J.E. (1996) The local repressor AcrR plays a modulating role in the regulation of acrAB genes of *Escherichia coli* by global stress signals. Mol Microbiol, 19, 101-112.

Madigan, M.T., Martinko, J.M. and Parker, J. (eds.). (2003) Brock Biology of Microorganisms, 10th edition. Pearson Education, Inc., Upper Saddle River, NJ 07458.

Martinon, F. and Tschopp, J. (2005) NLRs join TLRs as innate sensors of pathogens. Trends Immunol, 26, 447-454.

Mazmanian, S.K., Ton-That, H. and Schneewind, O. (2001) Sortase-catalysed anchoring of surface proteins to the cell wall of *Staphylococcus aureus*. Mol Microbiol, 40, 1049-1057.

Medrano, F.J., Gasset, M., Lopez-Zumel, C., Usobiaga, P., Garcia, J.L. and Menendez, M. (1996) Structural characterization of the unligated and choline-bound forms of the major pneumococcal autolysin LytA amidase. Conformational transitions induced by temperature. J Biol Chem, 271, 29152-29161.

Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics, 15, 211-218.

Mosavi, L.K. and Peng, Z.Y. (2003) Structure-based substitutions for increased solubility of a designed protein. Protein Eng, 16, 739-745.

Navarre, W.W. and Schneewind, O. (1999) Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol Mol Biol Rev, 63, 174-229.

Neuhaus, F.C. and Baddiley, J. (2003) A Continuum of Anionic Charge : Structures and Functions of D-Alanyl-Teichoic Acids in Gram-Positive Bacteria. Microbiology and Molecular Biology Reviews, 67, 686-723.

Nilsson, J., Stahl, S., Lundeberg, J., Uhlen, M. and Nygren, P.A. (1997) Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. Protein Expr Purif, 11, 1-16.

Nord, K., Gunneriusson, E., Uhl inverted question marken, M. and Nygren, P.A. (2000) Ligands selected from combinatorial libraries of protein A for use in affinity capture of apolipoprotein A-1M and taq DNA polymerase. J Biotechnol, 80, 45-54.

Ogunniyi, A.D., Giammarinaro, P. and Paton, J.C. (2002) The genes encoding virulence-associated proteins and the capsule of *Streptococcus pneumoniae* are upregulated and differentially expressed in vivo. Microbiology, 148, 2045-2053.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH--a hierarchic classification of protein domain structures. Structure, 5, 1093-1108.

Ortega, S., Garcia, J.L., Zazo, M., Varela, J., Munoz-Willery, I., Cuevas, P. and Gimenez-Gallego, G. (1992) Single-step purification on DEAE-sephacel of

recombinant polypeptides produced in *Escherichia coli*. Biotechnology (N Y), 10, 795-798.

Pedersen, A.K., Branner, S., Mortensen, S.B., Andersen, H.S., Klausen, K.M., Moller, K.B., Moller, N.P. and Iversen, L.F. (2004) Affinity purification of recombinant protein-tyrosine phosphatase 1B using a highly selective inhibitor. J Chromatogr B Analyt Technol Biomed Life Sci, 799, 1-8.

Pedone, E., Saviano, M., Rossi, M. and Bartolucci, S. (2001) A single point mutation (Glu85Arg) increases the stability of the thioredoxin from *Escherichia coli*. Protein Eng, 14, 255-260.

Persson, M., Bergstrand, M.G., Bulow, L. and Mosbach, K. (1988) Enzyme purification by genetically attached polycysteine and polyphenylalanine affinity tails. Anal Biochem, 172, 330-337.

Podvin, L., Reysset, G., Hubert, J. and Sebald, M. (1988) Presence of choline in teichoic acid of *Clostridium acetobutylicum* NI-4 and choline inhibition of autolytic function. J. Gen. Microbiol., 134, 1603-1609.

Poyart, C., Pellegrini, E., Marceau, M., Baptista, M., Jaubert, F., Lamy, M.C. and Trieu-Cuot, P. (2003) Attenuated virulence of *Streptococcus agalactiae* deficient in D-alanyl-lipoteichoic acid is due to an increased susceptibility to defensins and phagocytic cells. Mol Microbiol, 49, 1615-1625.

Ren, B., Szalai, A.J., Thomas, O., Hollingshead, S.K. and Briles, D.E. (2003) Both family 1 and family 2 PspA proteins can inhibit complement deposition and confer virulence to a capsular serotype 3 strain of *Streptococcus pneumoniae*. Infect Immun, 71, 75-85.

Rigden, D.J., Galperin, M.Y. and Jedrzejas, M.J. (2003) Analysis of structure and function of putative surface-exposed proteins encoded in the *Streptococcus pneumoniae* genome: a bioinformatics-based approach to vaccine and drug design. Crit Rev Biochem Mol Biol, 38, 143-168.

Robson, R.L. and Baddiley, J. (1977) Role of teichuronic acid in *Bacillus licheniformis*: defective autolysis due to deficiency of teichuronic acid in a novobiocin-resistant mutant. J Bacteriol, 129, 1051-1058.

Ronda, C., Garcia, J.L., Garcia, E., Sanchez-Puelles, J.M. and Lopez, R. (1987) Biological role of the pneumococcal amidase. Cloning of the lytA gene in *Streptococcus pneumoniae*. Eur J Biochem, 164, 621-624.

Ronda, C., Garcia, J.L. and Lopez, R. (1991) Teichoic acid choline esterase, a novel hydrolytic activity in *Streptococcus oralis*. FEMS Microbiol Lett, 64, 289-294.

Rosenow, C., Ryan, P., Weiser, J.N., Johnson, S., Fontan, P., Ortqvist, A. and Masure, H.R. (1997) Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of *Streptococcus pneumoniae*. Mol Microbiol, 25, 819-829.

Rost, B., Sander, C. and Schneider, R. (1994) PHD-an automatic mail server for protein secondary structure prediction. CABIOS, 10, 53-60.

Saiz, J.L., Lopez-Zumel, C., Monterroso, B., Varea, J., Arrondo, J.L., Iloro, I., Garcia, J.L., Laynez, J. and Menendez, M. (2002) Characterization of Ejl,

the cell-wall amidase coded by the pneumococcal bacteriophage Ej-1. Protein Sci, 11, 1788-1799.

Sanchez-Beato, A.R., Lopez, R. and Garcia, J.L. (1998) Molecular characterization of PcpA: a novel choline-binding protein of *Streptococcus pneumoniae*. FEMS Microbiol Lett, 164, 207-214.

Sanchez-Puelles, J.M., Ronda, C., Garcia, J.L., Garcia, P., Lopez, R. and Garcia, E. (1986) Searching for autolysin functions. Characterization of a pneumococcal mutant deleted in the lytA gene. Eur J Biochem, 158, 289-293.

Sanchez-Puelles, J.M., Sanz, J.M., Garcia, J.L. and Garcia, E. (1992) Immobilization and single-step purification of fusion proteins using DEAE-cellulose. Eur J Biochem, 203, 153-159.

Sanz, J.M., Diaz, E. and Garcia, J.L. (1992) Studies on the structure and function of the N-terminal domain of the pneumococcal murein hydrolases. Mol Microbiol, 6, 921-931.

Sanz, J.M., Garcia, P. and Garcia, J.L. (1996) Construction of a multifunctional pneumococcal murein hydrolase by module assembly. Eur J Biochem, 235, 601-605.

Sarvas, M., Harwood, C.R., Bron, S. and van Dijl, J.M. (2004) Post-translocational folding of secretory proteins in Gram-positive bacteria. Biochim Biophys Acta, 1694, 311-327.

Schleifer, K.H. and Kandler, O. (1972) Peptidoglycan types of bacterial cell walls and their taxonomic implications. Baceriol. Rev., 36, 407-477.

Schlieker, C., Bukau, B. and Mogk, A. (2002) Prevention and reversion of protein aggregation by molecular chaperones in the E. coli cytosol: implications for their applicability in biotechnology. J Biotechnol, 96, 13-21.

Schmidt, T.G. and Skerra, A. (1993) The random peptide library-assisted engineering of a C-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. Protein Eng, 6, 109-122.

Sheehan, M.M., Garcia, J.L., Lopez, R. and Garcia, P. (1997) The lytic enzyme of the pneumococcal phage Dp-1: a chimeric lysin of intergeneric origin. Mol Microbiol, 25, 717-725.

Smith, D.B. and Johnson, K.S. (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. Gene, 67, 31-40.

Smith, G.P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science, 228, 1315-1317.

Smith, J.C., Derbyshire, R.B., Cook, E., Dunthorne, L., Viney, J., Brewer, S.J., Sassenfeld, H.M. and Bell, L.D. (1984) Chemical synthesis and cloning of a poly(arginine)-coding gene fragment designed to aid polypeptide purification. Gene, 32, 321-327.

Smith, P.K., Krohn, R.I., Hermanson, G.T., Mallia, A.K., Gartner, F.H., Provenzano, M.D., Fujimoto, E.K., Goeke, N.M., Olson, B.J. and Klenk, D.C. (1985) Measurement of protein using bicinchoninic acid. Anal Biochem, 150, 76-85.

Song, J.K., Kim, M.K. and Rhee, J.S. (1999) Cloning and expression of the gene encoding phospholipase A1 from Serratia sp. MK1 in *Escherichia coli*. J Biotechnol, 72, 103-114.

Standish, A.J., Stroeher, U.H. and Paton, J.C. (2005) The two-component signal transduction system RR06/HK06 regulates expression of cbpA in *Streptococcus pneumoniae*. Proc Natl Acad Sci U S A, 102, 7701-7706.

Strub, C., Alies, C., Lougarre, A., Ladurantie, C., Czaplicki, J. and Fournier, D. (2004) Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. BMC Biochem, 5, 9.

Terpe, K. (2003) Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol, 60, 523-533.

Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., Heidelberg, J., DeBoy, R.T., Haft, D.H., Dodson, R.J., Durkin, A.S., Gwinn, M., Kolonay, J.F., Nelson, W.C., Peterson, J.D., Umayam, L.A., White, O., Salzberg, S.L., Lewis, M.R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A.M., Utterback, T.R., Hansen, C.L., McDonald, L.A., Feldblyum, T.V., Angiuoli, S., Dickinson, T., Hickey, E.K., Holt, I.E., Loftus, B.J., Yang, F., Smith, H.O., Venter, J.C., Dougherty, B.A., Morrison, D.A., Hollingshead, S.K. and Fraser, C.M. (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science, 293, 498-506.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Research, 22, 4673-4680.

Tomasz, A. (1967) Choline in the cell wall of a bacterium: novel type of polymer-linked choline in *Pneumococcus*. Science, 157, 694-697.

Tomasz, A. and Werner, T. (2000) The cell wall of *Streptococcus pneumoniae*. In Fischetti, V.A., Novick, R.P., Ferriti, J.J., Portnoy, D.A. and Rood, J.I. (eds.), Gram-Positive Pathogens. ASM Press, Washington.

Ton-That, H., Marraffini, L.A. and Schneewind, O. (2004) Protein sorting to the cell wall envelope of Gram-positive bacteria. Biochim Biophys Acta, 1694, 269-278.

Tuomanen, E. (1999) Molecular and cellular biology of pneumococcal infection. Curr Opin Microbiol, 2, 35-39.

Uhlen, M., Nilsson, B., Guss, B., Lindberg, M., Gatenbeck, S. and Philipson, L. (1983) Gene fusion vectors based on the gene for staphylococcal protein A. Gene, 23, 369-378.

Usobiaga, P., Medrano, F.J., Gasset, M., Garcia, J.L., Saiz, J.L., Rivas, G., Laynez, J. and Menendez, M. (1996) Structural organization of the major autolysin from *Streptococcus pneumoniae*. J Biol Chem, 271, 6832-6838.

Varea, J., Monterroso, B., Saiz, J.L., Lopez-Zumel, C., Garcia, J.L., Laynez, J., Garcia, P. and Menendez, M. (2004) Structural and thermodynamic characterization of Pal, a phage natural chimeric lysin active against pneumococci. J Biol Chem, 279, 43697-43707.

Varea, J., Saiz, J.L., Lopez-Zumel, C., Monterroso, B., Medrano, F.J., Arrondo, J.L., Iloro, I., Laynez, J., Garcia, J.L. and Menendez, M. (2000) Do sequence repeats play an equivalent role in the choline-binding module of pneumococcal LytA amidase? J Biol Chem, 275, 26842-26855.

Vollmer, W. and Tomasz, A. (2001) Identification of the teichoic acid phosphorylcholine esterase in *Streptococcus pneumoniae*. Mol Microbiol, 39, 1610-1622.

Ward, J.B. (1981) Teichoic and teichuronic acids: biosynthesis, assembly, and location. Microbiol Rev, 45, 211-243.

Weiser, J.N., Austrian, R., Sreenivasan, P.K. and Masure, H.R. (1994) Phase variation in pneumococcal opacity: relationship between colonial morphology and nasopharyngeal colonization. Infect Immun, 62, 2582-2589.

Weiser, J.N., Shchepetov, M. and Chong, S.T. (1997) Decoration of lipopolysaccharide with phosphorylcholine: a phase-variable characteristic of *Haemophilus influenzae*. Infect Immun, 65, 943-950.

Whiting, G.C. and Gillespie, S.H. (1996) Incorporation of choline into *Streptococcus pneumoniae* cell wall antigens: evidence for choline kinase activity. FEMS Microbiol Lett, 138, 141-145.

Wren, B.W. (1991) A family of clostridial and streptococcal ligand-binding proteins with conserved C-terminal repeat sequences. Mol Microbiol, 5, 797-803.

Xu, Y. and Xu, D. (2000) Protein threading using PROSPECT: design and evaluation. Proteins, 40, 343-354.

Yother, J., Handsome, G.L. and Briles, D.E. (1992) Truncated Forms of PspA That Are Secreted from *Streptococcus pneumoniae* and Their Use in Functional Studies and Cloning of the pspA Gene. journal of Bacteriology, 174, 610-618.

Yother, J. and White, J.M. (1994) Novel surface attachment mechanism of the *Streptococcus pneumoniae* protein PspA. J Bacteriol, 176, 2976-2985.

Zhang, J.R., Idanpaan-Heikkila, I., Fischer, W. and Tuomanen, E.I. (1999) Pneumococcal licD2 gene is involved in phosphorylcholine metabolism. Mol Microbiol, 31, 1477-1488.